Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

Cross-Level Mediation in School-Randomized Studies of Teacher Development: Experimental Design and Power

Ben Kelcey, Jessaca Spybrook, Nianbo Dong & Fangxing Bai

To cite this article: Ben Kelcey, Jessaca Spybrook, Nianbo Dong & Fangxing Bai (2020): Cross-Level Mediation in School-Randomized Studies of Teacher Development: Experimental Design and Power, Journal of Research on Educational Effectiveness, DOI: 10.1080/19345747.2020.1726540

To link to this article: https://doi.org/10.1080/19345747.2020.1726540

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
https://doi.org/10.1080/19345747.2020.1726540

Routledge
Taylor & Francis Group

METHODOLOGICAL STUDIES

Cross-Level Mediation in School-Randomized Studies of Teacher Development: Experimental Design and Power

Ben Kelcey[a], Jessaca Spybrook[b], Nianbo Dong[c] and Fangxing Bai[a]

[a]School of Education, University of Cincinnati, Cincinnati, Ohio, USA; [b]School of Education, Western Michigan University, Kalamazoo, Michigan, USA; [c]School of Education, University of North Carolina Chapel Hill, Chapel Hill, North Carolina, USA

ABSTRACT

Professional development for teachers is regarded as one of the principal pathways through which we can understand and cultivate effective teaching and improve student outcomes. A critical component of studies that seek to improve teaching through professional development is the detailed assessment of the intermediate teacher development processes that scaffold program content through three key types of outcomes—teacher knowledge, instruction, and student learning. Cross-level and sequential mediation strategies that probe and connect these processes and outcomes emerge as an important design consideration in these studies. We derive formulas that track the power with which school-randomized designs can detect professional development effects as they operate through a sequence of teacher-level mediators to affect student outcomes (e.g., school-randomized professional development studies). The results suggest that the sample sizes typically seen in well-planned experiments targeting total effects (e.g., 30–100 schools) can produce comparably high or disparately low levels of power for mediation effects—the similarity depends heavily on context and concomitant parameter values. The results are implemented in the *PowerUpR* package and in the *PowerUpR* Shiny application.

ARTICLE HISTORY
Received 6 June 2019
Revised 31 December 2019
Accepted 3 January 2020

KEYWORDS
Mediation; multilevel; teacher development

Professional development for teachers has been increasingly regarded as one of the principal pathways through which the field can understand and cultivate effective teaching and improve student outcomes (Correnti & Rowan, 2007). Recognition of professional development as a critical lever for change has spurred research and investments into teacher development programs (e.g., Institute of Education Sciences, 2019; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). A prominent issue in advancing teacher development at scale, however, has been developing theoretical and empirical models of effective professional development and delineating the channels and variables through which it operates (e.g., Scher & O'Reilly, 2009).

To address these limitations, recent literature has converged on a common theoretical infrastructure from which to systematically study teacher development processes

CONTACT Ben Kelcey ben.kelcey@gmail.com School of Education, University of Cincinnati, 2600 Clifton Ave., Cincinnati, Ohio 45220, USA.

Supplemental data for this article can be accessed at publisher's website.

(e.g., Desimone, 2009). Within this framework, teacher development is frequently conceptualized and tracked using three key types of outcomes: teacher knowledge, teaching quality, and student learning (Desimone, 2009). Systematic evaluation of each of these outcomes in succession formally develops and assesses theories of teacher change (e.g., how professional development improves knowledge, beliefs) and theories of instruction (e.g., how changes in knowledge or beliefs influence teacher instructional practice in ways that promote student outcomes; Desimone, 2009). In turn, these assessments provide a more comprehensive examination of professional development programs because they identify the impact of a program on student outcomes while detailing how and why the coordinated system of teacher development components comes to (or fails to) scaffold these effects through core intermediate teacher outcomes (i.e., indirect or mediation effect).

Recent efforts by scholars and funding agencies have invested widely in these types of studies that develop and delineate effective teaching and professional development strategies (Borko, 2004; Institute of Education Sciences, 2019). For instance, numerous studies have developed and examined programs that focus on teacher development, knowledge or instruction as it relates to student outcomes (e.g., Correnti & Rowan, 2007; Hill & Chin, 2018; Roth, Wilson, Taylor, Stuhlsatz, & Hvidsten, 2019). Moreover, funding agencies have commissioned scores of studies to develop, assess, and replicate the effects of programs operating through teachers.

Across many of these studies, one common design has been to assign schools rather than teachers to professional development and control conditions to accommodate the collaborative nature of many contemporary professional development programs and further minimize the potential for treatment diffusion (e.g., Rhoads, 2011; Roth et al., 2019; Spybrook, Shi, & Kelcey, 2016). The intersection of these design considerations and research objectives (i.e., school-level assignment, teacher-level intermediate outcome variables, and student-level outcomes) often gives rise to a cross-level mediation design—that is, designs that target how a school-level program (level three intervention) influences a teacher-level variable (level two mediator) in ways that improve a student-level outcome (level one outcome; Pituch, Murphy, & Tate, 2009). In this way, designing studies that delineate the teacher development processes that connect professional development with student learning emerge as an important consideration.

Although there is ample literature on the design of school-randomized studies for main and moderator effects, there is very less literature available to guide the design of professional development studies that seek to critically examine the coordinated system of relationships underpinning teacher development as it relates to student outcomes (e.g., Dong, Kelcey, & Spybrook, 2018; Raudenbush, 1997). In this study, we develop expressions to guide the design of cross-level mediation studies of school-wide professional development. We derive formulas that approximate the statistical power with which school-randomized designs can detect the indirect effects of professional development that pass through teacher-level mediators. We first consider 3-2-1 cross-level mediation designs that examine how professional development works through only one primary teacher variable such as teacher knowledge or instruction (Figure 1a). We then consider a more complete teacher development model by examining a 3-2-2-1 sequential mediation design in order to test both theories of teacher change (e.g., knowledge, beliefs) and theories of instruction

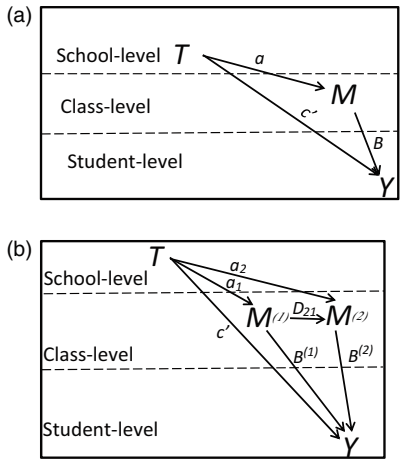

Figure 1. Conceptual diagrams of (a) 3-2-1 and (b) 3-2-2-1 mediation.

(e.g., improvements in instruction because of changes in knowledge or beliefs) that are commonly drawn on in research on teachers (Figure 1b; Desimone, 2009). The 3-2-2-1 design considers a sequence of teacher mediators—for example, how professional development (level three intervention) first improves teacher knowledge (level two proximal mediator), which then translates into changes in instruction (level two distal mediator), and ultimately results in improved student outcomes (level one outcome).

We first outline the indirect effects of professional development under the potential outcomes framework. We then develop expressions to address statistical power and sample size considerations in the 3-2-1 design and follow with the case of 3-2-2-1 design. In each of these sections, we probe the expressions governing power to get a sense of the relative influence of parameters and develop design strategies, assess the accuracy of our approximations with simulation and provide an example using *PowerUpR*. We end with a discussion.

Framework

Consider a 3-2-1 study in which researchers intend to track the indirect effect of a school-level professional development program on a student-level outcome as it works through a teacher-level intermediate variable such as teacher knowledge (Figure 1a; Roth et al., 2019). Our analyses track the difference between the potential student outcomes associated with the potential mediator values under the treatment condition and the control condition under the assumption that students are nested within only one primary teacher (e.g., no multiple membership or cross-classified structures). To track these differences, we begin with $Y_{ijk}(t_k)$ as the potential outcome for student i served by teacher j in school k when the school-level treatment (T_k) is set t_k (e.g., participation in professional development program or control condition) and $M_{jk}(t_k)$ as the potential mediator response for teacher j in school k when the school-level treatment (T_k) is set t_k (VanderWeele, Hong, Jones, & Brown, 2013). Within this context, the effects of participating in the professional development program (relative to a control condition) on a

teacher mediator (e.g., knowledge) and student outcome (e.g., achievement) can be expressed as $E[M_{jk}(1)-M_{jk}(0)]$ and $E[Y_{ijk}(1)-Y_{ijk}(0)]$ (VanderWeele et al., 2013).

Similarly, to examine mediation effects we describe the potential outcomes as a function of the treatment and mediator values. We use $Y_{ijk}(t_k, m_{jk})$ as the potential outcome for student i served by teacher j in school k when the school-level treatment (T_k) is set t_k and the (M_{jk}) is set m_{jk} (VanderWeele et al., 2013). The natural or total indirect effect is then (Muthén, 2011; VanderWeele, 2010; VanderWeele et al., 2013)

$$E\Big[Y_{ijk}(1, M_{jk}(1))-Y_{ijk}(1, M_{jk}(0))\Big] \tag{1}$$

The contrast of the above potential outcomes in Equation (1) describes, for example, the effect of a professional development program on student achievement that is due to changes in teacher knowledge (e.g., Carlisle, Kelcey, Rowan, & Phelps, 2011; Kelcey, 2011; Kelcey, Hill, & Chin, 2019).[1]

When individuals are the unit of assignment, the potential outcomes framework tracks effects by leaning on the single-level stable unit treatment value assumption (SUTVA). Individual-level SUTVA requires that the potential outcomes of a student do not depend on the program assigned to other students. The routine interactions among teachers within the same school and students served by the same teacher or school create a high propensity for this assumption to be violated (e.g., Hong & Raudenbush, 2008; Kelcey, 2011; VanderWeele et al., 2013). For this reason, educational research has regularly called upon on the assignment of intervention conditions to schools rather than individuals or teachers (e.g., Spybrook, Shi, & Kelcey, 2016). When the assignment to a program does not bring about the reassignment of students to new teachers or schools, designs using school-level assignment can adopt a weaker school-level version of SUTVA that requires only that the potential outcomes do not depend on the intervention condition assigned to other schools (i.e., school-level SUTVA; VanderWeele, 2008). In many studies in education, a reasonable approximation of this assumption is plausible because, for instance, schools are geographically disconnected and teachers can often be instructed to constrain relevant interactions during the study (Hong & Raudenbush, 2006).

The introduction of a teacher-level mediator (e.g., teacher knowledge), however, can complicate inferences because mediator values still vary among teachers within schools (VanderWeele, 2010). To help track indirect effects, we can further propose a teacher-level version of SUTVA that assumes the potential outcomes are not contingent upon the mediator values of other teachers inside or outside of a teacher's school (i.e., no interference among teachers/classrooms). However, even the teacher-level SUTVA may be unsustainable in some education situations because of, for example, the possibility of treatment diffusion—teachers within the same school intentionally or incidentally sharing key components of a program. For instance, contemporary professional development programs frequently encourage and leverage collaboration among teachers at the same school to promote/support the coordinated integration of the innovative methods into practice to achieve common aims (e.g., Desimone, 2009).

[1]More completely, this effect also incorporates changes in the relationship between teacher knowledge and student achievement produced by participation in the professional development program (see Kelcey, Dong, Spybrook, & Cox, 2017).

We can partially adapt SUTVA here by identifying the routes through which the teacher-level mediator values potentially contribute to the potential outcomes (e.g., Hong & Raudenbush, 2006; VanderWeele, 2010; VanderWeele et al., 2013). If schools are sufficiently independent such that potential outcomes plausibly depend on only the mediator values of teachers inside the same school but not on the mediator values of teachers from other schools, then we can re-express the potential outcomes in light of the mediator values of other teachers in the same school. Let $m_{.k}$ be $\{m_{jk}\}_{j=1}^{n_{2k}}$, the vector of mediator values of the n_{2k} teachers inside school k and m_{-jk} identify the vector of mediator values for all teachers other than teacher j in school k.

To simplify inferences in this situation, prior research has constrained the additional role of the mediator on the potential outcomes to work through a scalar function of the mediator values of other teachers in a school (Hong & Raudenbush, 2006; VanderWeele, 2010; VanderWeele et al., 2013). Let $f(m_{-jk})$ be a scalar function of the other teachers' mediator values in a school (e.g., school-level mean) with realization f. We can re-state the potential outcome for student i served by teacher j in school k as a function of the treatment (t_k), the mediator m_{jk}, and a scalar function(s) of the mediator values of other teachers at a school $f(m_{-jk})$ as follows:

$$Y_{ijk}(t_k, m_{jk,}\, m_{-jk}) = Y_{ijk}(t_k, m_{jk,} f(m_{-jk})) \tag{2}$$

Within this framework, we use the potential outcomes to describe the movement of effects from the school-level program to the student-level outcome via a teacher-level mediator (e.g., Pituch & Stapleton, 2012; VanderWeele, 2010; VanderWeele et al., 2013). More specifically, our analyses target the natural or total indirect effect (TIE) as described through the contrast of the following potential outcomes (VanderWeele et al., 2013)

$$\text{TIE} = E\Big[Y_{ijk}(1, M_{jk}(1), f(M_{-jk}(1))) - Y_{ijk}(1, M_{jk}(0), f(M_{-jk}(0)))\Big] \tag{3}$$

The first potential outcome details the response of student i served by teacher j in school k when the mediator values of all teachers in a school correspond to the value they would take when assigned to the professional development program. The second potential outcome then corresponds to a student's response under the treatment condition but when the mediator values of all teachers in a school are set to the value they would take under the control condition. Because the differences between these potential outcomes simultaneously modulate the mediator values of all teachers in a school, it captures how the indirect effects flow through changes in the value of an individual teacher mediator (e.g., knowledge of individual teachers) and through a scalar function (e.g., average) of other teachers' mediator values at a school (e.g., the collective knowledge base of all teachers; contextual effects).

Within this structure, the indirect effect is identified under three primary conditions (for more details, see VanderWeele et al., 2013). The first assumption is the consistency of the observed responses such that they correspond to the potential responses. The second assumption is that there are no downstream variables that are influenced by program participation and subsequently confound the outcome-mediator relationship (e.g., Avin, Shpitser, & Pearl, 2005). The third assumption is that the potential mediator and outcome responses are independent once we condition on additional covariates (i.e., sequential ignorability). That is when we condition on additional variables such as

student-level covariates X_{ijk}, their teacher-/class- ($g(X_{.jk})$), and school-level functions ($h(X_{.k})$), teacher-/class-level covariates W_{jk}, their school-level functions ($q(W_{.k})$), and school-level covariates $\boldsymbol{Z}_k$, this assumption necessitates the following equations:

$$\{M_{jk}(t), f(M_{-jk}(t))\} \coprod T_k | X_{ijk}, W_{jk}, Z_k, g(X_{.jk}), h(X_{.k}), q(W_{.k}) \tag{4}$$

$$Y_{ijk}(t, m, f) \coprod \{M_{jk}(t), f(M_{-jk}(t))\} | T_k, X_{ijk}, W_{jk}, Z_k, g(X_{.jk}), h(X_{.k}), q(W_{.k}) \tag{5}$$

with $\coprod$ denoting conditional independence. In our setting, Equation (4) is supported by the random assignment of schools to conditions. However, Equation (5) becomes reasonable if and only if we are able to control for all variables that confound the mediator-outcome pathway. Further embedded in this framework is that sequential ignorability is supported only when (a) the teacher potential mediator responses and functions of those responses are conditionally independent of the program assignment and when (b) the potential student outcomes are conditionally independent of both the teacher mediators and functions of those responses given treatment condition, student-, teacher-/class-, and school-level covariates and their functions. Below we condition on covariates to approximate these assumptions and develop power analysis formulas; however, we caution that sequential ignorability is a critical assumption that should be carefully considered in practice.

3-2-1 Mediation

Prior literature has proposed and investigated multiple approaches to estimate multilevel mediation effects under a variety of different designs and assumptions (e.g., Bauer, Preacher, & Gil, 2006; Hong & Nomi, 2012; Krull & MacKinnon, 2001; Qin & Hong, 2017). One common approach has been to approximate the indirect effect using multilevel models (Pituch et al., 2009; Pituch & Stapleton, 2012; VanderWeele et al., 2013; VanderWeele & Vanseteelandt, 2009; Zhang, Zyphur, & Preacher, 2009).

Drawing on prior literature, we structure the mediation model as a three-level random intercept model that conditions on school-, teacher-/class-, and student-level covariates, and their respective cluster means (Pituch & Stapleton, 2012; Raudenbush & Bryk, 2002; VanderWeele, 2010; VanderWeele et al., 2013; Zhang et al., 2009). Because the most common cluster-level scalar function of lower-level variables is the mean, we detail our analyses principally within this purview (Pituch & Stapleton, 2012; Raudenbush & Bryk, 2002; VanderWeele et al., 2013). We use group-mean centered variables because this is common in the literature and is convenient for separating out indirect effects across levels (e.g., Brincks et al., 2017; Pituch & Stapleton, 2012; Zhang et al., 2009). Alternative approaches such as using the original data (no centering) or grand-mean centering can be transformed to yield the same parameter estimates under random intercept formulations (Brincks et al., 2017; Enders & Tofighi, 2007; Kreft, De Leeuw, & Aiken, 1995).

Assuming additivity and constant effects, we specify the mediator model as follows:

$$\begin{array}{ll} M_{jk} = \pi_{0k} + \pi_1(\bar{X}_{jk} - \bar{X}_k) + \pi_2(W_{jk} - \bar{W}_k) + \pi_3 Q_{jk} + \varepsilon_{jk}^M & \varepsilon_{jk}^M \sim N(0, \sigma_{M|}^2) \\ \pi_{0k} = \zeta_{00} + aT_k + \zeta_{01}\bar{X}_k + \zeta_{02}\bar{W}_k + \zeta_{03}Z_k + u_{0k}^M & u_{0k}^M \sim N(0, \tau_{M|}^2) \end{array} \tag{7}$$

For this mediator model, M_{jk} is used as the original mediator value for teacher j in school k, $\bar{X}_{jk}$ as a teacher-level mean of a student-level covariate (with π_1 as its path coefficient) and $\bar{X}_k$ as its school-level mean aggregate (with ζ_{01} as its path coefficient), W_{jk} as a teacher-level covariate (π_2 as its path coefficient) and $\bar{W}_k$ as its school-level mean aggregate (ζ_{02} as its path coefficient), Q_{jk} as a teacher-level variable that varies only among teachers within schools (no variation among schools) with π_3 as its path coefficient, T_k as the intervention assignment coded as $\pm 1/2$ with associated path coefficient a, Z_k as a school-level covariate (ζ_{03} as its path coefficient), and ε_{jk}^M as the teacher-level error term, and u_{0k}^M as the school-level random effects. If we constrain the model so that there are no intervention-by-mediator interactions, the outcome model is as follows:

$$\begin{gathered} Y_{ijk} = \beta_{0jk} + \beta_1(X_{ijk} - \bar{X}_{jk}) + \beta_2 V_{ijk} + \varepsilon_{ijk}^Y \qquad \varepsilon_{ijk}^Y \sim N(0, \sigma_{Y|}^2) \\ \beta_{0jk} = \gamma_{00k} + b_2(M_{jk} - \bar{M}_k) + \gamma_{01}(\bar{X}_{jk} - \bar{X}_k) + \gamma_{02}(W_{jk} - \bar{W}_k) + \gamma_{03} Q_{jk} + u_{0jk}^Y \qquad u_{0jk}^Y \sim N(0, \tau_{Y|}^2) \\ \gamma_{00k} = \zeta_0 + B\bar{M}_k + c' T_k + \xi_1 \bar{X}_k + \xi_2 \bar{W}_k + \xi_3 Z_k + \nu_{00k}^Y \qquad \nu_{00k}^Y \sim N(0, \upsilon_{Y|}^2) \end{gathered} \tag{8}$$

We additionally use Y_{ijk} as the outcome for student i served by teacher j in school k, X_{ijk} as a student-level covariate (with β_1 as its path coefficient), V_{ijk} as a student-level covariate that only varies among students within teachers/classrooms (no variation among teachers or schools) with β_2 as its path coefficient, $M_{jk} - \bar{M}_k$ as the school-centered teacher-level mediator with coefficient b_2, $\bar{M}_k$ as the mean of the mediator in school k with path coefficient B, c' as the intervention-outcome conditional path coefficient, γ as coefficients for teacher-level covariates, ξ as coefficients for school-level covariates, and υ_{00k}^Y, u_{0jk}^Y and ε_{ijk}^Y as the school, teacher, and student error terms. Alternatively, we could additionally consider the potential for an interaction between the intervention and mediator in our model (e.g., adding a $T\bar{M}_k$ term). For simplicity, we outline the derivations and results under the assumption that researchers do not anticipate an interaction. However, including such an interaction in the design phase is a straightforward extension (Kelcey et al., 2017).

Using group-mean centering, the coefficient (B) attached to the school-level mediator mean captures the total conditional association between the mediator and the outcome. In this representation, b_2 captures the association between the teacher-level mediator and the outcome while the difference of the coefficients ($B - b_2$) captures the unique additional association between the school-level mediator (aggregate) and the outcome (Kreft et al., 1995; Raudenbush & Bryk, 2002). As a result, past literature has leveraged the B coefficient to summarize the total (teacher plus school) conditional association of the mediator and outcome.

Given this formulation, past research (Pituch & Stapleton, 2012; VanderWeele, 2008, 2010; VanderWeele & Vansteelandt, 2009) has demonstrated that the total indirect effect (TIE) can be estimated using TIE = aB. In terms of our professional development example, the indirect effect describes the total improvement in student achievement that accumulates as a result of improvements in both individual teacher knowledge and in the school-wide improvement in teacher knowledge produced through participation in the professional development program.

Error Variance

To develop expressions to guide the planning of three-level school-randomized studies probing mediation, we first develop approximations to the error variance of each path coefficient and then the total indirect effect under maximum likelihood estimation. Under the assumption of sequential ignorability and unrelated errors across the mediator and outcome models, the covariance of the parameters comprising the indirect effect is asymptotically zero under linear models (i.e., $\sigma_{a,B} = 0$; Allison, 1995). As a result, the variance of the indirect effect is (Allison, 1995; Bollen, 1987; MacKinnon, Lockwood, Brown, Wang, & Hoffman, 2007):

$$\sigma^2_{aB} = \sigma^2_a B^2 + a^2\sigma^2_B + \sigma^2_a\sigma^2_B \tag{9}$$

where σ^2_a and σ^2_B are the error variances of the respective path coefficients.

The asymptotic distribution of each path coefficient under maximum likelihood estimation is normal with a mean as the true coefficient value and covariance equal to the inverse of the information matrix. When data has been collected, the covariance matrix of the coefficients is often estimated by the inverse of the observed information matrix (Raudenbush & Bryk, 2002). However, this matrix is unavailable when planning a study. To overcome this constraint, we draw on the expected information matrix based on the functions of model parameters and common summary statistics.

Our mediator model above allows for individual-level covariates (V_{ijk}, X_{ijk}), teacher-level averages ($\bar{X}_{jk}$) and school-level averages ($\bar{X}_k$), teacher-level covariates (Q_{jk}, W_{jk}) and school-level averages ($\bar{W}_k$), and a school-level covariate (Z_k). We standardize the variables (save intervention) to have a mean of zero and variance of one. In matrix form, the mediation model is as follows:

$$\begin{array}{ll} M_k = W_k\pi_k + \varepsilon_k & \varepsilon_k \sim N(0, \sigma^2_{M|}I) \\ \pi_k = Z_k\zeta + u_k & u_k \sim N(0, T_{M|}) \end{array} \tag{10}$$

with k in 1 to n_3 and W_k capturing all teacher-level variables and Z_k capturing all school-level variables (Raudenbush & Bryk, 2002). The first expression of Equation (10) captures the differences among teachers such that M_k is an n_{2k} by one vector of mediator values, W_k is an n_{2k} by four matrix consisting of a column of a unit vector, Q values, and centered X and W values, π_k is a four by one vector of unknown path coefficients, I is an n_2 by n_2 identity matrix, and $\boldsymbol{\varepsilon}_k$ is an n_{2k} by one vector of errors with mean vector zero and a variance-covariance matrix with variances as $\sigma^2_{M|}$ and covariances of zero. At the school-level, Z_k can take the form of a four by four matrix of school predictors (T, $\bar{X}_k$, $\bar{W}_k$, and Z_k) but it is reduced to a one by four matrix under a random intercept model. Similarly, u_k can be reduced to a one by one matrix consisting of only the random intercept with $T_{M|}$ as the covariance matrix. Last, ζ is a four by one vector of fixed effects (representing a, ζ_{01}, ζ_{02}, ζ_{03} in Equation (7)).

With maximum likelihood estimation and an equal number of teachers per school and an equal number of students per teacher, the conditional dispersion of π_k given Z_k is $Var(\hat{\pi}_k) = T + V_k = \Delta_k$ with V_k the error variance matrix such that $V_k = \sigma^2_{M|}(W_k^T W_k)^{-1}$. With an equal number of teachers inside each school, the dispersions of the $\hat{\pi}_k$'s are constant such that $\Delta_k = \Delta = T_{M|} + \sigma^2_{M|}(W^T W)^{-1}$. Matrix algebra reduces the second element pertaining to the intercept $\sigma^2_{M|}(W^T W)^{-1}_{11} = \sigma^2_{M|}/n_2$ (Kelcey et al.,

2017). $T_{M|}$ can also be reduced to a scalar, $\tau^2_{M|}$, in our model because the elements beyond the first in the u_k matrix are zero for each school. As a result, we get the following equation:

$$\Delta_k = \Delta = T_{M|} + \sigma^2_{M|}(W^T W)^{-1}_{11} = \tau^2_{M|} + \sigma^2_{M|}/n_2 \tag{11}$$

Using maximum-likelihood, the dispersion matrix of the school-level coefficients on the random intercept is $V_{\hat{\zeta}} = Var(\hat{\zeta}) = \Delta(\sum Z_k^T Z_k)^{-1}$. Evaluating this expression shows that the diagonal element representing the intervention path coefficient can be obtained as follows:

$$\sigma^2_a = \Delta\frac{1}{n_3 p(1-p)} = \frac{\tau^2_{M|} + \sigma^2_{M|}/n_2}{n_3 p(1-p)} = \frac{\tau^2_M(1-R^2_{M^{L3}}) + (1-R^2_{M^{L2}})\sigma^2_M/n_2}{n_3 p(1-p)} \tag{12}$$

Here, τ^2_M and σ^2_M represent the unconditional school- and teacher-level variances of the mediator, p represents the proportion of schools exposed to the program, $R^2_{M^{L3}}$ and $R^2_{M^{L2}}$ represent the school- and teacher-level mediator variance explained by predictors in the mediator model (Equations (7)) and n_3 and n_2 represent the school- and teacher-level sample sizes.

B Path

We apply a similar analysis to trace to the error variance of the B path in the outcome model. The results demonstrate that the error variance of the B path can be tracked as:

$$\sigma^2_B = \frac{\upsilon^2_{Y|} + \tau^2_{Y|}/n_2 + \sigma^2_{Y|}/(n_2 n_1)}{n_3(\tau^2_{M|} + \sigma^2_{M|}/n_2)} = \frac{\upsilon^2_Y(1-R^2_{Y^{L3}}) + \tau^2_Y(1-R^2_{Y^{L2}})/n_2 + (1-R^2_{Y^{L1}})\sigma^2_Y/(n_2 n_1)}{n_3(\tau^2_M(1-R^2_{M^{L3}}) + (1-R^2_{M^{L2}})\sigma^2_M/n_2)} \tag{13}$$

Our result further introduces υ^2_Y, τ^2_Y, and σ^2_Y as the unconditional variances for the outcome at the school-, teacher- and student-levels. We summarize the contribution of the predictors using $R^2_{Y^{L3}}$, $R^2_{Y^{L2}}$ and $R^2_{Y^{L1}}$ as the school-, teacher-, and student-level outcome variances explained by other predictors in the outcome model (Equation (8)).

Literature developing similar expressions has typically used a first-order approximation of the variance by dropping the final product of the path error variances term in Equation (9) because it is comparatively small (e.g., $\sigma^2_a\sigma^2_B \approx 0$; Sobel, 1982). Under the first-order approximation, the resulting error variance of the indirect effect becomes:

$$\begin{aligned}\sigma^2_{aB} = &\left(\frac{\tau^2_M(1-R^2_{M^{L3}}) + \sigma^2_M(1-R^2_{M^{L2}})/n_2}{n_3 p(1-p)}\right)B^2 \\ &+ a^2\left(\frac{\upsilon^2_Y(1-R^2_{Y^{L3}}) + \tau^2_Y(1-R^2_{Y^{L2}})/n_2 + \sigma^2_Y(1-R^2_{Y^{L1}})/(n_2 n_1)}{n_3(\tau^2_M(1-R^2_{M^{L3}}) + \sigma^2_M(1-R^2_{M^{L2}})/n_2)}\right)\end{aligned} \tag{14}$$

We can also standardize this expression by fixing the total outcome and mediator variances to be one such that $\tau^2_M + \sigma^2_M = \rho_M + (1-\rho_M) = 1$ and $\upsilon^2_Y + \tau^2_Y + \sigma^2_Y = \rho_{Y_3} + \rho_{Y_2} + (1-\rho_{Y_3}-\rho_{Y_2}) = 1$. As a result, the error variance can be expressed as follows:

$$\begin{aligned}\sigma^2_{aB} = &\left(\frac{\rho_M(1-R^2_{M^{L3}}) + (1-R^2_{M^{L2}})(1-\rho_M)/n_2}{n_3 p(1-p)}\right)B^2 \\ &+ a^2\left(\frac{\rho_{Y_3}(1-R^2_{Y^{L3}}) + \rho_{Y_2}(1-R^2_{Y^{L2}})/n_2 + (1-R^2_{Y^{L1}})(1-\rho_{Y_3}-\rho_{Y_2})/(n_2 n_1)}{n_3(\rho_M(1-R^2_{M^{L3}}) + (1-R^2_{M^{L2}})(1-\rho_M)/n_2)}\right)\end{aligned} \tag{15}$$

With such standardization the path coefficients can also be interpreted alongside conventional scales—for example, the a and c' paths can be interpreted along a standardized mean difference scale and the B path can be interpreted along a standardized regression coefficient scale.

Error Variance Properties

A longstanding challenge for education studies has been that school-randomized designs typically demand larger sample sizes to achieve reasonable levels of power (e.g., Raudenbush, Martinez, & Spybrook, 2007). For this reason, prior literature has developed design principles and strategies to understand and potentially improve the efficiency of studies (e.g., Raudenbush et al., 2007). To uncover potential design principles and strategies for mediation, we probed the properties of the error variance with respect to its governing parameters. Similar investigations of, for example, the main effect error variance have suggested a range of design principles and strategies that can be used to guide design and improve efficiency (e.g., Kelcey & Shen, 2019; Raudenbush, 1997). Prior research on study design with regard to main effects has uncovered simple principles such as the dominant role of the cluster relative to the individual sample size in determining the error variance and strategies such as covariance adjustment to improve efficiency (e.g., Raudenbush et al., 2007). For this reason, we probed the behavior of the error variance expressions by examining how the error variance of the indirect effect changes with modulations of each parameter holding other parameters constant (e.g., Usami, 2019). Using derivatives of the error variance, we organized our analysis along three types of parameters: (a) the sample size (n_3, n_2, n_1), (b) parameters associated with outcome variance ($\rho_{Y_3}, \rho_{Y_2}, R^2_{Y^{L3}}, R^2_{Y^{L2}}, R^2_{Y^{L1}}$), and (c) parameters associated with the mediator variance ($\rho_M, R^2_{M^{L3}}, R^2_{M^{L2}}$).

Sample Size

Consistent with other types of multilevel analyses, the error variance had an inverse relationship with sample size at each level with the school-level sample size playing the most important role, followed by the teacher-level sample size and then the student-level sample size (see example below for illustration).

Outcome Intraclass Correlation Coefficients

Next, we considered the intraclass correlation coefficients for the outcome (ρ_{Y_3}andρ_{Y_2}). The first derivative of the error variance of the indirect effect in terms of the school-level intraclass correlation coefficient (ρ_{Y_3}) was:

$$\frac{\partial \sigma^2_{aB}}{\partial \rho_{Y_3}} = a^2\left(\frac{(1-R^2_{Y^{L3}})-(1-R^2_{Y^{L1}})/n_1 n_2}{n_3(\rho_M(1-R^2_{M^{L3}})+(1-\rho_M)(1-R^2_{M^{L2}})/n_2)}\right) \tag{16}$$

Analysis of this derivative indicates that the relationship between the school-level intraclass correlation coefficient and the error variance of the indirect effect will typically be positive—that is, larger values of the school-level intraclass correlation coefficient will produce larger error variances. Analysis of the teacher-level intraclass correlation

coefficient (ρ_{Y_2}) returns a similar pattern. The positive nature of the relationship parallels a well-known finding in power analyses for main effects—larger intraclass correlation coefficients will typically increase the uncertainty of the estimate and reduce the power of a study holding other factors equal. However, this relationship is slightly more complicated than with total effects. The potential exception to the positive relationship for indirect effects arises when both the number of students per school ($n_1 n_2$) is very small (e.g., $n_1 = 2$, $n_2 = 2$, $n_1 n_2 = 4$) and when the outcome variance explained at the school-level substantially exceeds that of the student-level. In such scenarios, the error variance is materially unaffected by increases in the school-level intraclass correlation coefficient because that clustering is explained by covariates while concurrently implying smaller individual-level variance. The implication is that by and large increases in the school-level outcome variance will yield larger standard errors for estimates of the indirect effect.

We also considered the relative influence of the school and teacher intraclass correlation coefficients. Taking the difference of the respective derivatives demonstrates once again that the school-level intraclass correlation coefficient will typically be more influential in shaping the error variance (and power) than that of the teacher-level. However, in atypical situations when the number of teachers is small (e.g., $n_2 = 2$) and when $R^2_{Y^{L2}}$ is very small relative to $R^2_{Y^{L3}}$, ρ_{Y_2} can be occasionally more influential than ρ_{Y_3} because increases in school-level clustering are effectively explained away by a covariate so that clustering at the class-level dominates the error variance.

Outcome Variance Explained

We next examined the contributions of the variance in the outcome explained by predictors at each level. At the school level, the resulting first derivative was always negative:

$$\frac{d\sigma^2_{aB}}{dR^2_{Y^{L3}}} = a^2\left(-\frac{\rho_{Y^{L3}}}{n_3(\rho_M(1-R^2_{M^{L3}}) + (1-R^2_{M^{L2}})(1-\rho_M)/n_2)}\right) \leq 0 \tag{17}$$

Intuitively, increases in the variance explained at the school-level produce decreases in the error variance of estimates of the indirect effect. These results further extend to the outcome variance explained at other levels—increases in the variance explained at the teacher-level and student-level also returned smaller standard errors holding other values constant. Overall, the relationships between the error variance and the parameters associated with the outcome were intuitive and paralleled the relationships typically seen in the analysis of the main effects.

Mediator Intraclass Correlation Coefficients

Analysis of parameters associated with the mediator demonstrated much more complex relationships. We first examined the error variance in terms of the mediator intraclass correlation coefficient. The results demonstrated the relationship was heavily contingent upon the values other parameters took. For instance, when the proportion of variance explained in the mediator by predictors was smaller than that explained in the outcome, there tended to be a positive relationship—increases in the mediator intraclass correlation coefficient were paired with increases in the error variance of the indirect effect. However, when the proportion of

variance explained in the mediator by predictors was similar or higher than that explained in the outcome, there tended to be a negative relationship—increases in the mediator intraclass correlation were paired with decreases in the error variance.

Mediator Variance Explained

Similarly complicated results were obtained for the relationship between the error variance and the variance explained in the mediator by predictors. Increases in the mediator variance explained by predictors can yield smaller or larger standard errors depending on the values of other parameters. Conceptually, although explaining variation in the mediator decreases the error variance associated with the a path coefficient, it can inflate the error variance associated with the B path coefficient. As a result, the relative value of explaining variance in the mediator depends on two competing forces within the error variance expression—the reduction in the contribution of the uncertainty in the a path against the inflation of uncertainty in the B path. Prior literature in single-level studies has documented this complex relationship (e.g., Beasley, 2014). We further probe these complexities by unpacking it as a function of the path coefficients below.

Path coefficient formulation

Although the error variance expressions developed above outline how the uncertainty of the estimated indirect effect changes as a function of parameters, the variance explained components (e.g., $R^2_{Y^{L3}}$) conflate the contributions of the focal variables (e.g., treatment, mediator) with those of the covariates. For instance, the $R^2_{Y^{L3}}$ term in Equation (13) represents the collective variance explained by the mediator, treatment, and covariates. When planning a study, however, researchers often conceptually segregate the contributions of variables such that they describe the focal paths (i.e., a, B, c') using the anticipated effect sizes but the contribution of the covariates using their collective variance explained (R^2). In order to accommodate this approach, we restructure the error variances by decomposing them in terms of the focal path coefficients and the variance explained by covariates. In doing so, we can potentially draw on prior empirical literature to specify plausible values of the variance explained by covariates (e.g., Hedges & Hedberg, 2007; Konstantopoulos, 2012) while specifying the anticipated effect sizes for the focal paths. For the school-level variance explained in the outcome ($R^2_{Y^{L3}}$) we have:

$$R^2_{Y^{L3}} = R^2_{Y^{L3}_{\vec{Z}}} + \frac{p(1-p)(aB+c\prime)^2}{\upsilon^2_Y} + \frac{\tau^2_M + \sigma^2_M/n_2}{\upsilon^2_Y} B^2\left(1 - \frac{p(1-p)a^2}{\tau^2_M} - R^2_{M^{L3}_{\vec{Z}}}\right) \tag{18}$$

Here, we use $\vec{Z}$ to represent the vector of covariates presented in the outcome model. The first term on the right-hand side ($R^2_{Y^{L3}_{\vec{Z}}}$) describes the school-level squared multiple correlations between the outcome and the covariates ($R^2_{Y^{L3}_{\vec{Z}}}$) whereas the last term ($R^2_{M^{L3}_{\vec{Z}}}$) describes the same but for the mediator at the school level. In contrast to the previous expressions (e.g., in Equation (13), the $R^2_{Y^{L3}_{\vec{Z}}}$ parameter align with the variance explained often reported in empirical studies of design parameter values (e.g., Kelcey & Phelps, 2013a).

For the outcome variance explained by predictors at the teacher-level ($R^2_{Y^{L2}}$), we have the following equation:

$$R^2_{Y^{L2}} = R^2_{Y^{L2}_{\vec{Z}}} + \left(\frac{\sigma^2_M}{\tau^2_Y}\right) b^2_2 (1 - R^2_{M^{L2}_{\vec{Z}}}) \tag{19}$$

with $R^2_{Y^{L2}_{\vec{Z}}}$ and $R^2_{M^{L2}_{\vec{Z}}}$ capturing the total teacher-level outcome and mediator variance explained by covariates. For the outcome variance explained at the student-level ($R^2_{Y^{L1}}$), we simply have $R^2_{Y^{L1}} = R^2_{Y^{L1}_{\vec{Z}}}$, with $R^2_{Y^{L1}_{\vec{Z}}}$ as the variance explained by student-level covariates.

We can take a parallel approach for the mediator. The total teacher-level variance explained for the mediator ($R^2_{M^{L2}}$) is simply the variance explained by the covariates (i.e., $R^2_{M^{L2}} = R^2_{M^{L2}_{\vec{Z}}}$) whereas the total school-level variance explained ($R^2_{M^{L3}}$) can be expanded as

$$R^2_{M^{L3}} = R^2_{M^{L3}_{\vec{Z}}} + \frac{p(1-p)a^2}{\tau^2_M} \tag{20}$$

Hypothesis Tests & Power

We next extended three test statistics and their associated power functions using the previous developments. We considered two different asymptotic-based tests and one resampling-based test–each of which can be employed in the planning stages before data collection has begun: (a) Sobel test, (b) the joint test, and (c) the Monte Carlo interval test.

Sobel Test

The Sobel test statistic contrasts the ratio of the estimated indirect effect on its standard error outlined before (Sobel, 1982):

$$z^{Sobel}_{aB} = aB/\sigma_{aB} \tag{21}$$

with σ_{aB} as the square root of the before-referenced error variance of the indirect effect. Inferences under the Sobel test are typically drawn by contrasting the test statistic with a normal distribution on the basis that the maximum likelihood estimates are asymptotically normal with mean equal to the true indirect effect and variance equal to Equation (14) before (Sobel, 1982). The corresponding power of the test can be approximated using:

$$P(|z^{Sobel}_{aB}| > z_{critical}) = 1 - \Phi(z_{critical} - z^{Sobel}_{aB}) + \Phi(-z_{critical} - z^{Sobel}_{aB}) \tag{22}$$

where z^{Sobel}_{aB} is the Sobel test statistic outlined above, Φ represents the cumulative normal density function, and with $z_{critical}$ is the critical value (e.g., 1.96) from the normal distribution corresponding to a particular type one error rate.

The comparison of the Sobel test statistic to a normal distribution can be reasonable under-sample sizes that are large. With small to moderate sample sizes, however, the normal distribution can serve as a poor referent distribution because the sampling distribution of the indirect effect can be heavily skewed (Kisbu-Sakarya, MacKinnon, & Miočević, 2014).

Joint Test

A simple but high-powered alternative to the Sobel test is the joint test (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). The joint test develops inferences using sub-tests that target the path coefficients that compose an indirect effect—a sub-test for the intervention-mediator path coefficient and, separately, a sub-test for the mediator-outcome path coefficient. Under the joint test, the null hypothesis of no indirect effect is rejected only when both sub-tests are rejected. Although the joint test is limited in that it does not provide confidence intervals, past research has shown that the joint test performs well and returns results similar to resampling-based tests including those based on bootstrap methods in single-level contexts (Hayes & Scharkow, 2013).

For the program-mediator (a) and mediator-outcome (B) paths, we use the test statistics:

$t_a = a/\sigma_a$ and $t_B = B/\sigma_B$(23) where σ_a and σ_a are the standard errors above. For each of these sub-tests we use a referent t-distribution with degrees of freedom equal to n_3 – $C-$ 1 where C is the number of school-level predictors in the mediator model for the a path and the outcome model for the B path (Raudenbush & Bryk, 2002; Kenny & Judd, 2014).

The power of two-sided tests to detect the indirect effect (i.e., both paths concurrently nonzero) is the product of the power to detect the intervention-mediator path and the power to detect the corresponding mediator-outcome path:

$$\begin{aligned} P(|t_a|>t_{critical}\&|t_B|>t_{critical}) = (1-t(t_{critical}-t_a) + t(-t_{critical}-t_a)) \\ \times (1-t(t_{critical}-t_B) + t(-t_{critical}-t_B)) \end{aligned} \quad (24)$$

where t is the appropriate cumulative t density function and $t_{critical}$ is the corresponding critical value for the appropriate n_3 – $C-$ 1 degrees of freedom.

Monte Carlo Interval Test

Last, we consider the resampling-based Monte Carlo interval test (Preacher & Selig, 2012). In this test, plausible values are drawn for the intervention-mediator and the mediator-outcome path coefficients from normal distributions centered at their estimated values $(\hat{a}, \hat{B})$ with variances set to their expected error variances based on the formulas developed above. By drawing plausible values, we can approximate the sampling distribution of the indirect effect using the product of a^* and B^* (Preacher & Selig, 2012):

$$\begin{pmatrix} a^* \\ B^* \end{pmatrix} \sim t_{n_3-C-1}\left[\begin{pmatrix} \hat{a} \\ \hat{B} \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2_{\hat{a}} & \hat{\sigma}_{\hat{a},\hat{B}} \\ \hat{\sigma}_{\hat{a},\hat{B}} & \hat{\sigma}^2_{\hat{B}} \end{pmatrix}\right] \quad (25)$$

Inferences regarding the indirect effect are then made on the basis of whether the simulated asymmetric confidence intervals exclude zero. The statistical power of the test can then be evaluated using the proportion of asymmetric confidence intervals (e.g., 95%) that exclude zero. Like the joint test, literature has demonstrated that the Monte Carlo interval test performs comparable to bootstrap-based methods with the advantages of being much less computationally intensive and being feasible in the design phase when data has not yet been collected.

Table 1. Simulated and predicted power for 3-2-1 mediation.

	n_3	n_2	n_1	a	B	ρ_Y^{L3}	ρ_Y^{L2}	ρ_M	$R^2_{Y_2^{L3}}$	$R^2_{Y_2^{L2}}$	$R^2_{Y_2^{L1}}$	$R^2_{M_2^{L3}}$	$R^2_{M_2^{L2}}$	$\widehat{Sobel}$	Sobel	$\widehat{Joint}$	Joint	$\widehat{MC}$	MC
1	30	4	20	0.49	0.30	0.20	0.15	0.26	0.38	0.41	0.02	0.17	0.07	0.31	0.19	0.30	0.35	0.31	0.33
2	40	4	20	0.50	0.30	0.20	0.15	0.27	0.38	0.41	0.02	0.16	0.07	0.43	0.35	0.51	0.55	0.53	0.53
3	50	4	20	0.50	0.30	0.19	0.15	0.27	0.38	0.41	0.02	0.16	0.07	0.52	0.49	0.66	0.67	0.67	0.66
4	60	4	20	0.51	0.30	0.19	0.15	0.27	0.38	0.41	0.02	0.16	0.07	0.62	0.67	0.79	0.81	0.82	0.79
5	80	4	20	0.50	0.30	0.19	0.15	0.27	0.38	0.41	0.02	0.15	0.07	0.76	0.85	0.91	0.92	0.92	0.91
6	100	4	20	0.50	0.30	0.19	0.15	0.27	0.38	0.41	0.02	0.16	0.07	0.85	0.94	0.97	0.96	0.98	0.96
7	200	4	20	0.50	0.30	0.20	0.15	0.27	0.38	0.41	0.02	0.16	0.07	0.99	1.00	1.00	1.00	1.00	1.00
8	30	2	20	0.49	0.30	0.11	0.10	0.13	0.27	0.26	0.01	0.07	0.03	0.14	0.06	0.13	0.10	0.15	0.12
9	40	2	20	0.51	0.30	0.11	0.10	0.12	0.28	0.24	0.01	0.10	0.03	0.19	0.10	0.19	0.13	0.21	0.16
10	50	2	20	0.49	0.30	0.11	0.10	0.12	0.27	0.25	0.01	0.10	0.02	0.22	0.18	0.22	0.17	0.25	0.24
11	60	2	20	0.52	0.30	0.11	0.10	0.11	0.28	0.25	0.01	0.10	0.03	0.28	0.21	0.29	0.24	0.31	0.26
12	80	2	20	0.51	0.30	0.11	0.10	0.11	0.28	0.24	0.01	0.13	0.02	0.36	0.30	0.38	0.33	0.39	0.37
13	100	2	20	0.50	0.30	0.11	0.10	0.10	0.28	0.25	0.01	0.11	0.02	0.42	0.44	0.44	0.44	0.44	0.46
14	200	2	20	0.49	0.30	0.11	0.10	0.10	0.28	0.25	0.01	0.14	0.02	0.71	0.70	0.74	0.71	0.75	0.71
15	30	4	10	0.50	0.30	0.28	0.23	0.31	0.70	0.74	0.09	0.42	0.24	0.31	0.16	0.30	0.34	0.30	0.31
16	50	4	10	0.50	0.30	0.29	0.23	0.30	0.71	0.74	0.09	0.43	0.25	0.52	0.49	0.66	0.69	0.67	0.67
17	80	4	10	0.49	0.30	0.29	0.23	0.31	0.71	0.74	0.09	0.43	0.24	0.74	0.84	0.90	0.91	0.92	0.91
18	100	4	10	0.50	0.30	0.29	0.23	0.31	0.71	0.74	0.09	0.43	0.24	0.84	0.95	0.97	0.98	0.96	0.97
19	150	4	10	0.50	0.30	0.29	0.23	0.30	0.71	0.74	0.09	0.43	0.24	0.96	1.00	1.00	1.00	1.00	1.00
20	30	10	20	0.52	0.29	0.08	0.10	0.09	0.36	0.24	0.01	0.16	0.02	0.30	0.18	0.30	0.32	0.32	0.30
21	40	10	20	0.50	0.30	0.08	0.10	0.09	0.36	0.25	0.01	0.17	0.02	0.41	0.30	0.48	0.50	0.51	0.50
22	50	10	20	0.50	0.30	0.08	0.10	0.09	0.36	0.24	0.01	0.16	0.02	0.50	0.48	0.62	0.65	0.62	0.64
23	70	10	20	0.50	0.30	0.08	0.10	0.09	0.35	0.24	0.01	0.16	0.02	0.66	0.73	0.80	0.81	0.81	0.81
24	80	10	20	0.50	0.30	0.08	0.10	0.09	0.36	0.24	0.01	0.16	0.02	0.72	0.81	0.85	0.87	0.86	0.87
25	30	10	20	0.30	0.20	0.07	0.10	0.08	0.34	0.24	0.01	0.18	0.02	0.14	0.03	0.07	0.09	0.06	0.07
26	50	10	20	0.29	0.20	0.07	0.10	0.08	0.36	0.25	0.01	0.18	0.02	0.21	0.08	0.19	0.21	0.19	0.19
27	70	10	20	0.30	0.20	0.07	0.10	0.08	0.35	0.24	0.01	0.18	0.02	0.30	0.19	0.34	0.36	0.34	0.36
28	100	10	20	0.29	0.20	0.07	0.10	0.08	0.35	0.24	0.01	0.19	0.02	0.41	0.34	0.49	0.50	0.47	0.49
29	200	10	20	0.30	0.20	0.07	0.10	0.08	0.35	0.24	0.01	0.18	0.02	0.73	0.79	0.84	0.84	0.84	0.84
30	30	10	20	0.81	0.50	0.13	0.09	0.12	0.29	0.25	0.01	0.12	0.02	0.64	0.65	0.76	0.77	0.79	0.77
31	50	10	20	0.81	0.50	0.13	0.09	0.12	0.29	0.25	0.01	0.12	0.02	0.88	0.96	0.96	0.97	0.97	0.97
32	70	10	20	0.80	0.50	0.13	0.09	0.12	0.30	0.24	0.01	0.13	0.02	0.97	0.99	0.99	0.99	0.99	0.99
33	30	10	20	0.79	0.30	0.08	0.10	0.12	0.33	0.24	0.01	0.12	0.02	0.48	0.42	0.58	0.61	0.60	0.59
34	50	10	20	0.80	0.30	0.08	0.10	0.12	0.33	0.25	0.01	0.11	0.02	0.74	0.84	0.91	0.92	0.93	0.91
35	100	10	20	0.81	0.30	0.08	0.10	0.12	0.34	0.24	0.01	0.12	0.02	0.97	1.00	1.00	1.00	1.00	1.00
36	30	10	20	0.31	0.50	0.11	0.10	0.08	0.33	0.24	0.01	0.17	0.02	0.18	0.11	0.17	0.16	0.19	0.17
37	50	10	20	0.30	0.50	0.11	0.10	0.08	0.33	0.25	0.01	0.18	0.02	0.28	0.27	0.28	0.30	0.31	0.32
38	80	10	20	0.31	0.50	0.11	0.10	0.08	0.34	0.24	0.01	0.19	0.02	0.44	0.44	0.46	0.46	0.46	0.48
39	100	10	20	0.30	0.50	0.11	0.10	0.08	0.33	0.25	0.01	0.17	0.02	0.53	0.54	0.54	0.56	0.57	0.56
													Ave	0.53	0.51	0.59	0.59	0.60	0.60

Simulation

To probe the accuracy and utility of our expressions, we used Monte Carlo simulations to contrast the simulated type one error and power rates for detecting the indirect effect with our formula-based power predictions across 1000 simulated data sets for each condition. We generated data using the mediation models detailed in Equations (7) and (8) and considered school sample sizes spanning 30–200 along with a variety of conditions for other parameters that are outlined in Table 1.

The results of the simulation are detailed in Tables 1 and 2 under the following column labels: (a) $\widehat{Sobel}$ (predicted power or type one error rate based on Sobel expressions above), (b) *Sobel* (simulated power or type one error rate for Sobel test), (c) $\widehat{Joint}$ (predicted power or type one error rate based on Joint expressions above), (d) *Joint* (simulated power or type one error rate based on Joint test), (e) $\widehat{MC}$ (predicted power or type one error rate based on the Monte Carlo interval test developed above), (f) *MC* (simulated power or type one error rate based on the Monte Carlo interval test). There were consistently close correspondences between the simulated power level and our predicted power level for the joint test and Monte Carlo interval test and to a lesser extent the Sobel test (Table 1). The most powerful test was the Monte Carlo interval test and across conditions the average absolute difference

Table 2. Simulated and predicted type one error rates for 3-2-1 mediation.

	n_3	n_2	n_1	a	B	ρ_Y^{L3}	ρ_Y^{L2}	ρ_M	$R^2_{Y^{L3}_2}$	$R^2_{Y^{L2}_2}$	$R^2_{Y^{L1}_2}$	$R^2_{M^{L3}_2}$	$R^2_{M^{L2}_2}$	$\widehat{Sobel}$	Sobel	$\widehat{Joint}$	Joint	$\widehat{MC}$	MC
40	30	4	20	0	0.3	0.29	0.23	0.29	0.73	0.73	0.09	0.51	0.24	0.05	0	0.03	0.01	0.01	0.01
41	50	4	20	0	0.3	0.28	0.23	0.28	0.73	0.74	0.09	0.5	0.24	0.05	0.01	0.05	0.02	0.02	0.02
42	60	4	20	0	0.3	0.28	0.23	0.28	0.73	0.73	0.09	0.5	0.24	0.05	0.01	0.05	0.02	0.02	0.02
43	80	4	20	0	0.3	0.28	0.23	0.28	0.73	0.73	0.09	0.5	0.23	0.05	0.01	0.05	0.03	0.03	0.03
44	100	4	20	0	0.3	0.28	0.23	0.27	0.73	0.73	0.09	0.5	0.24	0.05	0.02	0.05	0.02	0.03	0.03
45	200	4	20	0	0.3	0.28	0.23	0.28	0.73	0.73	0.09	0.5	0.24	0.05	0.02	0.05	0.03	0.03	0.03
46	30	4	20	0.5	0	0.2	0.23	0.31	0.67	0.71	0.09	0.43	0.24	0.05	0.01	0.02	0.02	0.01	0.02
47	50	4	20	0.5	0	0.2	0.23	0.31	0.66	0.71	0.09	0.42	0.23	0.05	0.01	0.04	0.02	0.01	0.02
48	70	4	20	0.5	0	0.2	0.23	0.3	0.66	0.71	0.09	0.42	0.24	0.05	0.01	0.04	0.02	0.03	0.02
49	100	4	20	0.5	0	0.2	0.23	0.31	0.66	0.71	0.09	0.44	0.24	0.05	0.01	0.05	0.03	0.02	0.03
50	200	4	20	0.5	0	0.2	0.23	0.31	0.66	0.71	0.09	0.43	0.24	0.05	0.01	0.05	0.02	0.03	0.02
51	30	4	20	0	0	0.2	0.22	0.27	0.67	0.71	0.09	0.5	0.24	0.05	0	0	0	0	0
52	50	4	20	0	0	0.19	0.23	0.28	0.65	0.71	0.09	0.49	0.23	0.05	0	0	0	0	0
53	100	4	20	0	0	0.2	0.23	0.28	0.66	0.71	0.09	0.5	0.24	0.05	0	0	0	0	0
													Ave	0.05	0.01	0.03	0.02	0.02	0.02

between the predicted and simulated power was just 0.01. The joint test was nearly as powerful as the Monte Carlo interval test and the discrepancy between the simulated and predicted power averaged only 0.02. The Sobel test was the least powerful and averaged discrepancies of about 0.06—however, the predicted power of the Sobel test tended to experience larger discrepancies for smaller school-level sample sizes than larger school-level sample sizes (see Table 1). For type one error rates (Table 2), the results suggested that all tests maintained rates lower than the adopted 5% level—our formulations accurately predicted these less than nominal rates for the Monte Carlo interval test and the joint test but the predictions were less precise for the Sobel test because of its reliance on the normal distribution for the indirect effect as its reference point.

Recent research probing indirect effects has emphasized the stringent assumptions required to causally identify indirect effects (e.g., VanderWeele et al., 2013). As noted earlier, two key assumptions supporting the interpretation of indirect effects are sequential ignorability and no treatment by mediator interaction. Our simulations additionally probed the sensitivity of our results to violations of no mediator-outcome confounding variable assumption and the no treatment-by-mediator interaction assumption we invoked in our formulation. We generated two additional types of simulations: (a) a mediator and outcome that was a function of an unobserved variable u (with coefficient U for both responses ranging from 0.1 to 0.5) and (b) an outcome that was a function of an interaction between the mediator and treatment (with coefficient $\underline{\underline{\Delta}}$ ranging from 0.1 to 0.3). We then predicted the power rates using the aforementioned expressions that omit the consideration of the unobserved variable and interaction. The results are outlined in Supplementary Tables A1 and A2 by the magnitude of the coefficients. The results suggest that discrepancies between the predicted and true power were proportional to the magnitude of the coefficients. Under most instances, the formulas still provide a reasonable approximation of power (e.g., within 5% of the true value). However, when the magnitude of the coefficient for the unobserved variable or for the interaction is large (e.g., 0.5) the discrepancies can rise to 10% or 15%. Collectively, the results suggest that consideration of the assumptions for a given study is critical because violations have the potential to undermine the validity and utility of the formulas and present misleading estimates of power.

Example

Consider an example study that intends to use a school-randomized design to probe the impact of an early elementary professional development program in reading (intervention) on student reading achievement (outcome) as it operates through improvements in teacher knowledge (mediator). Based on prior literature, presume that we anticipate that the outcome (achievement) variance attributable to schools is 0.15, to teachers is 0.15 and to students is 0.70 while the mediator (teacher knowledge) variance attributable to schools is 0.20 and teachers is 0.80 (e.g., Hedges & Hedberg, 2007; Kelcey & Phelps, 2013a; Kelcey, Shen, & Spybrook, 2016; Westine et al., in review). Further, assume that we intend to secure several covariates in our data collection efforts including pretests for students and teachers. We anticipate that together the covariates will predict roughly 25% of the variance at each level for the mediator and outcome (e.g., Hedges & Hedberg, 2007; Kelcey & Phelps, 2013b; Westine et al., in review). Based on previous research and pilot studies, presume we would like to design a study to detect a total or main effect as small as 0.25 (standardized difference between conditions) with the impact of the professional development program on teacher knowledge (mediator) as 0.50 (standardized difference between conditions) and the conditional association between teacher knowledge and student achievement (standardized regression coefficient scale) as 0.30 such that the indirect effect is $0.5 \times 0.3 = 0.15$ (e.g., Kowalski et al., 2019; Phelps, Kelcey, Jones, & Liu, 2016; Scher & O'Reilly, 2009). Further, assume that the direct effect of the program on student achievement is $c' = 0.10$ such that a simple decomposition of the total effect (c) is $0.25 = 0.5 \times 0.3 + 0.1$ ($c = aB + c'$). That is, let $a = 0.5$, $B = 0.30$ $b_2 = 0.1$, $c' = 0.10$ $R^2_{Y_z^{L3}} = R^2_{Y_z^{L2}} = R^2_{Y_z^{L1}} = R^2_{M_z^{L3}} = R^2_{M_z^{L2}} = 0.25$. If we plan to sample 20 students per teacher (n_1), four teachers per school (n_2), how many schools do we need to yield approximately 80% power?

To identify the requisite school-level sample size, we carry out the analyses using the *R* package *PowerUpR* and the corresponding *Shiny* application *PowerUpR Shiny* (https://poweruprshiny.shinyapps.io/v104). Figure 2a outlines the resulting power functions for the example by detailing the power for the Sobel (long-dash curve), joint (short-dash curve), and Monte Carlo interval tests (solid line curve) for the indirect effect as a function of school-level sample size (n_3). Our application indicates that under the Monte Carlo interval and joint tests roughly 54 schools would yield a power level close to 0.80. For the Sobel test we would need nearly 80 schools. Although not shown here, power analysis for the main effect under the same conditions indicated that 80 schools would be needed to achieve power of approximately 0.80. Hence if the study was adequately powered to detect the main effect of treatment, it would also be powered to detect the indirect effect, particularly under the Monte Carlo and joint tests.

Careful design practice suggests the consideration of a plausible range of parameter values to gauge the sensitivity of the results to fallible estimates (e.g., Cox & Kelcey, 2019). For instance, prior research has suggested that statistical power can be sensitive to the assumed values of the intraclass correlation and variance explained coefficients (e.g., Hedges & Hedberg, 2007). In Supplementary Figure A1, we probe this sensitivity for our illustrative analysis by considering a plausible range of values for these parameters (i.e., intraclass correlation coefficient values between 0.10 and 0.25 and variance explained values between 0.10 and 0.50). The results suggested that the requisite school sample size was fairly insensitive to errors in the variance explained parameters but

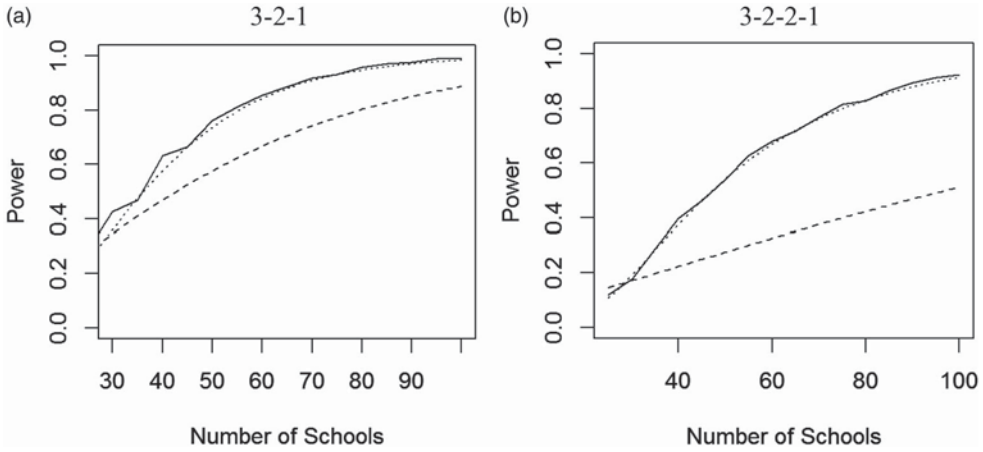

Figure 2. Power as a function of the number of schools for (a) 3-2-1 and (b) 3-2-2-1 under the Sobel test (long dash), Joint test (short dash) and Monte Carlo interval test (solid line).

sensitive to errors in the intraclass correlation coefficient parameters (see Supplementary Figure A1). In the extreme case, intraclass correlation coefficients of about 0.25 would drive the requisite sample size to 76 schools.

3-2-2-1 Sequential Mediation

We next extend our analyses to the case of a 3-2-2-1 sequential mediation design to probe the progression of teacher development using a sequence of intermediate teacher variables. As outlined earlier, a key application of this design arises when schools are assigned to participate in a professional development program or comparison condition and the program is expected to first change teacher knowledge, which in turn is expected to improve the type or quality of instruction teachers offer, which then is expected to advance student outcomes (e.g., Desimone, 2009). One recent example of this type of study was a cluster-randomized trial designed to assess the impact of a Content-Focused Coaching professional development program on reading achievement for fourth and fifth grade students (Matsumura, Garnier, & Spybrook, 2013). The Content-Focused Coaching program drew on instructional coaches to provide teachers training to improve their literacy content knowledge. The expectation was that the new content knowledge for teachers would increase teachers' capacity to facilitate high-quality text discussions during their instruction which would lead to increased reading achievement for students. The resulting study can be described as a type of 3-2-2-1 sequential mediation design (Figure 1b).

Given multiple mediators, there are multiple types of indirect effects that can be of interest (Daniel, Stavola, Cousens, & Vansteelandt, 2015). Our focus on the progression of teacher development processes targets the three-path or sequential indirect effect transmitted successively through the first (e.g., knowledge) and second (e.g., instruction) mediators to the outcome (tracing the $a_1 D_{21} B^{(2)}$ paths in Figure 1b; Taylor, MacKinnon, & Tein, 2008). Using previous notation, our targeted sequential indirect effect is:

$$\begin{aligned} &E[Y_{ijk}(1, M_{jk}^{(1)}(1), f(M_{-jk}^{(1)}(1)), M_{jk}^{(2)}(1, M_{jk}^{(1)}(1)), f(M_{-jk}^{(2)}(1, f(M_{-jk}^{(1)}(1)))))\\ &-Y_{ijk}(1, M_{jk}^{(1)}(1), f(M_{-jk}^{(1)}(1)), M_{jk}^{(2)}(1, M_{jk}^{(1)}(0)), f(M_{-jk}^{(2)}(1, f(M_{-jk}^{(1)}(0)))))] \end{aligned} \tag{26}$$

with $M_{jk}^{(1)}$ and $M_{jk}^{(2)}$ as the first (e.g., knowledge) and second (e.g., instruction) mediators and $f(M_{-jk}^{(1)})$ and $f(M_{-jk}^{(2)})$ as scalar functions of the vector of other teachers' first and second mediator values within a school.

Causal interpretation of mediation analyses with multiple sequential mediators is notoriously complex and identification of indirect effects and path-specific effects requires a list of formidable assumptions that additionally require a set of parametric modeling suppositions to make estimation tractable (e.g., Daniel et al., 2015; Steen, Loeys, Moerkerke, & Vansteelandt, 2017; Taylor et al., 2008). For the indirect effect we target, these assumptions include each of those in the 3-2-1 case along with their multivariate extensions. However, identification also requires more complex versions of sequential ignorability and no interactions (e.g., between the treatment and either mediator or between mediators). For a comprehensive analysis of the identifiability of effects see, for example, Daniel et al. (2015) and Steen et al. (2017).

To make study planning tractable here, we accept these assumptions and operationalize them using parametric linear structural equation models to identify path–specific effects (e.g., Taylor et al., 2008). However, we again caution that the veracity of this approach and the causal interpretation of parameter estimates are deeply dependent on the validity of assumptions and the assumed parametric form. Close scrutiny of each of the assumptions and careful development of models is a critical precondition to the planning and use of our subsequent developments regarding power and sample sizes.

Expanding our previous 3-2-1 models, we now consider a larger system of equations to describe the relationships and theory of action among the treatment, mediators, and outcome (Taylor et al., 2008). Assuming additivity and constant effects, we specify the model for the first or proximal mediator (e.g., teacher knowledge) as:

$$\begin{aligned} M_{jk}^{(1)} &= \pi_{0k}^{(1)} + \pi_1^{(1)}(\bar{X}_{jk} - \bar{X}_k) + \pi_2^{(1)}(W_{jk} - \bar{W}_k) + \pi_3^{(1)} Q_{jk} + \varepsilon_{jk}^{M^{(1)}} \qquad \varepsilon_{jk}^{M^{(1)}} \sim N(0, \sigma^2_{M^{(1)}_{|}})\\ \pi_{0k}^{(1)} &= \zeta_{00}^{(1)} + a_1 T_k + \zeta_{01}^{(1)} \bar{X}_k + \zeta_{02}^{(1)} \bar{W}_k + \zeta_{03}^{(1)} Z_k + u_{0k}^{M^{(1)}} \qquad u_{0k}^{M^{(1)}} \sim N(0, \tau^2_{M^{(1)}_{|}}) \end{aligned} \tag{27}$$

Notation remains unchanged from the 3-2-1 analysis except that we now introduce superscripts to denote the model focus on the first or proximal mediator (e.g., knowledge) and a subscript on the impact coefficient (a_1) to quantify the direct effect of the program on the first mediator.

For the second or distal mediator (e.g., instruction), we specify the model as follows:

$$\begin{aligned} M_{jk}^{(2)} &= \pi_{0k}^{(2)} + d_{21}(M_{jk}^{(1)} - \bar{M}_k^{(1)}) + \pi_1^{(2)}(\bar{X}_{jk} - \bar{X}_k) + \pi_2^{(2)}(W_{jk} - \bar{W}_k) + \pi_3^{(2)} Q_{jk} + \varepsilon_{jk}^{M^{(2)}} \qquad \varepsilon_{jk}^{M^{(2)}} \sim N(0, \sigma^2_{M^{(2)}_{|}})\\ \pi_{0k}^{(2)} &= \zeta_{00}^{(2)} + D_{21} \bar{M}_k^{(1)} + a_2 T_k + \zeta_{01}^{(2)} \bar{X}_k + \zeta_{02}^{(2)} \bar{W}_k + \zeta_{03}^{(2)} Z_k + u_{0k}^{M^{(2)}} \qquad u_{0k}^{M^{(2)}} \sim N(0, \tau^2_{M^{(2)}_{|}}) \end{aligned} \tag{28}$$

We expand previous notation to introduce $M_{jk}^{(2)}$ as the value of the second mediator (e.g., instruction) for teacher j in school k, D_{21} as the total association between the mediators with d_{21} as the teacher-level component and a_2 as the direct effect of the program on the second mediator.

Last, we specify the outcome model (e.g., student achievement) as:

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + \beta_1(X_{ijk} - \bar{X}_{jk}) + \beta_2 V_{ijk} + \varepsilon^Y_{ijk} \qquad \varepsilon^Y_{ijk} \sim N(0, \sigma^2_{Y|}) \\ \beta_{0jk} &= \gamma_{00k} + b^{(2)}_2(M^{(2)}_{jk} - \bar{M}^{(2)}_k) + b^{(1)}_2(M^{(1)}_{jk} - \bar{M}^{(1)}_k) + \gamma_{01}(\bar{X}_{jk} - \bar{X}_k) \\ &\quad + \gamma_{02}(W_{jk} - \bar{W}_k) + \gamma_{03} Q_{jk} + u^Y_{0jk} \qquad u^Y_{0jk} \sim N(0, \tau^2_{Y|}) \\ \gamma_{00k} &= \zeta_0 + B^{(2)}\bar{M}^{(2)}_k + B^{(1)}\bar{M}^{(1)}_k + c' T_k + \xi_1 \bar{X}_k + \xi_2 \bar{W}_k + \xi_3 Z_k + \nu^Y_{00k} \qquad \nu^Y_{00k} \sim N(0, \upsilon^2_{Y|}) \end{aligned} \tag{29}$$

Once again, we expand the previous notation and introduce $B^{(2)}$ to capture the total conditional association between the outcome and the second mediator with $b^{(2)}_2$ as the teacher-level component of that relationship and $B^{(1)}$ to capture the total conditional association between the outcome and the first mediator with $b^{(1)}_2$ as the teacher-level component of that relationship.

Under no interactions among the treatment, first, and second mediators, the sequential indirect effect can be estimated as $a_1 D_{21} B^{(2)}$ (Taylor et al., 2008). In terms of our professional development example, this sequential indirect effect describes how participation in a professional development program induces a change in teachers' knowledge in ways that yield higher-quality instruction and subsequently manifest as improvements in student achievement.

Error Variance

Similar to the 3-2-1 case, a good approximation to the variance of the sequential indirect effect is (e.g., Taylor et al., 2008):

$$\sigma^2_{a_1 D_{21} B^{(2)}} = a_1^2 D_{21}^2 \sigma^2_{B^{(2)}} + a_1^2 (B^{(2)})^2 \sigma^2_{D_{21}} + D_{21}^2 (B^{(2)})^2 \sigma^2_{a_1} \tag{30}$$

where $\sigma^2_{a_1}$, $\sigma^2_{D_{21}}$ and $\sigma^2_{B^{(2)}}$ are the error variances of the respective path coefficients. In turn, analytic expressions approximating each of the path error variances can be obtained using the expected information. For the a_1 path, the error variance ($\sigma^2_{a_1}$) is:

$$\sigma^2_{a_1} = \left(\frac{\tau^2_{M^{(1)}_{|}} + \sigma^2_{M^{(1)}_{|}}/n_2}{n_3 p(1-p)} \right) = \left(\frac{\tau^2_{M^{(1)}}(1 - R^2_{M^{(1)}_{L3}}) + \left(1 - R^2_{M^{(1)}_{L2}}\right)\sigma^2_{M^{(1)}}/n_2}{n_3 p(1-p)} \right) \tag{31}$$

The notation is adapted from Equation (12) in a straightforward manner. The variance explained in the first mediator at the school- ($R^2_{M^{(1)}_{L3}}$) and teacher-level ($R^2_{M^{(1)}_{L2}}$) can be unpacked as follows:

$$R^2_{M^{(1)}_{L3}} = R^2_{M^{(1)}_{L3\vec{Z}}} + \frac{p(1-p)a_1^2}{\tau^2_{M^{(1)}}} \text{ and } R^2_{M^{(1)}_{L2}} = R^2_{M^{(1)}_{L2\vec{Z}}} \tag{32}$$

with $R^2_{M^{(1)}_{L3\vec{Z}}}$ and $R^2_{M^{(1)}_{L2\vec{Z}}}$ as the variance explained by covariates at the respective level.

Similarly, the error variance for D_{21} path ($\sigma^2_{D_{21}}$) reduces to:

$$\sigma^2_{D_{21}} = \left(\frac{\tau^2_{M^{(2)}_{|}} + \sigma^2_{M^{(2)}_{|}}/n_2}{n_3 \sigma^2_T (1 - R^2_T)} \right) = \left(\frac{\tau^2_{M^{(2)}}(1 - R^2_{M^{(2)}_{L3}}) + (1 - R^2_{M^{(2)}_{L2}})\sigma^2_{M^{(2)}}/n_2}{n_3(\tau^2_{M^{(1)}}(1 - R^2_{M^{(1)}_{L3}}) + (1 - R^2_{M^{(1)}_{L2}})\sigma^2_{M^{(1)}}/n_2)} \right) \tag{33}$$

Here $R^2_{M^{(2)}_{L3}}$ and $R^2_{M^{(2)}_{L2}}$ describe the variance explained in the second mediator at the school- and teacher-level by covariates, the treatment, and the first mediator. They can

be further evaluated as:

$$R^2_{M^{(2)}_{L3}} = R^2_{M^{(2)}_{L3\vec{Z}}} + \frac{p(1-p)(a_1 D_{21} + a_2)^2}{\tau^2_{M^{(2)}}} + \frac{\tau^2_{M^{(1)}} + \sigma^2_{M^{(1)}}/n_2}{\tau^2_{M^{(2)}}} D^2_{21}(1-R^2_{M^{(1)}_{L3}})$$
$$R^2_{M^{(2)}_{L2}} = R^2_{M^{(2)}_{L2\vec{Z}}} + \left(\frac{\sigma^2_{M^{(1)}}}{\sigma^2_{M^{(2)}}}\right) d^2_{21}(1-R^2_{M^{(1)}_{L2\vec{Z}}}) \quad (34ab)$$

Last, the error variance for $B^{(2)}$ path ($\sigma^2_{D_{21}}$) is

$$\sigma^2_{B^{(2)}} = \frac{\upsilon^2_{Y|} + \tau^2_{Y|}/n_2 + \sigma^2_{Y|}/(n_2 n_1)}{n_3(\tau^2_{M^{(2)}_{|}} + \sigma^2_{M^{(2)}_{|}}/n_2)} = \frac{\upsilon^2_Y(1-R^2_{Y^{L3}}) + \tau^2_Y(1-R^2_{Y^{L2}})/n_2 + (1-R^2_{Y^{L1}})\sigma^2_Y/(n_2 n_1)}{n_3(\tau^2_{M^{(2)}}(1-R^2_{M^{(2)}_{L3}}) + (1-R^2_{M^{(2)}_{L2}})\sigma^2_{M^{(2)}}/n_2)} \quad (35)$$

where the outcome variance explained terms (R^2_Y) include the contributions of both mediators (e.g., $R^2_{Y^{L3}}$ includes the contributions of $\bar{M}^{(2)}$, $\bar{M}^{(1)}$, T, and covariates) and the terms describing the variance explained in the second mediator ($R^2_{M^{(2)}}$) include the contributions of the first mediator (e.g., $R^2_{M^{(2)}_{L3}}$ includes the contributions of $\bar{M}^{(1)}$, T, and covariates). The variances explained can be approximated as follows:

$$R^2_{Y^{L3}} = R^2_{Y^{L3}_{\vec{Z}}} + R^2_{Y^{L3}_{T}} + R^2_{Y^{L3}_{M^{(1)}|T\vec{z}}} + R^2_{Y^{L3}_{M^{(2)}|M^{(1)}T\vec{Z}}} \quad (36)$$

with

$$R^2_{Y^{L3}_{T}} = \frac{p(1-p)(a_1 B^{(1)} + a_2 B^{(2)} + a_1 D_{21} B^{(2)} + c\prime)^2}{\upsilon^2_Y}$$
$$R^2_{Y^{L3}_{M^{(1)}|TZ}} = \frac{\tau^2_{M^{(1)}} + \sigma^2_{M^{(1)}}/n_2}{\upsilon^2_Y}(D_{21}B^{(2)} + B^{(1)})^2(1-R^2_{M^{(1)}_{L3}}) \quad (37)$$
$$R^2_{Y^{L3}_{M^{(2)}|M^{(1)}TZ}} = \frac{\tau^2_{M^{(2)}} + \sigma^2_{M^{(2)}}/n_2}{\upsilon^2_Y}(B^{(2)})^2\left(1-R^2_{M^{(2)}_{L3}}\left(\frac{\tau^2_{M^{(2)}}}{\tau^2_{M^{(2)}} + \sigma^2_{M^{(2)}}/n_2}\right)\right)$$

At the teacher-level, the variance explained expands to:

$$R^2_{Y^{L2}} = R^2_{Y^{L2}_{\vec{Z}}} + \frac{\sigma^2_{M^{(1)}}}{\tau^2_Y}(d_{21}b^{(2)}_2 + b^{(1)}_2)^2(1-R^2_{M^{(1)}_{L2}}) + \frac{\sigma^2_{M^{(2)}}}{\tau^2_Y}(b^{(2)}_2)^2(1-R^2_{M^{(2)}_{L2}}) \quad (38)$$

Hypothesis Tests and Power

Sobel Test

The power of the Sobel test for the 3-2-2-1 indirect effect can be approximated using:

$$P(|z^{Sobel}_{a_1 D_{21} B^{(2)}}| > z_{critical}) = 1-\Phi(z_{critical} - z^{Sobel}_{a_1 D_{21} B^{(2)}}) + \Phi(-z_{critical} - z^{Sobel}_{a_1 D_{21} B^{(2)}}) \quad (39)$$

with $z^{Sobel}_{a_1 D_{21} B^{(2)}} = a_1 D_{21} B^{(2)}/\sigma_{a_1 D_{21} B^{(2)}}$ and $\sigma_{a_1 D_{21} B^{(2)}}$ as the square root of the error variance.

Joint Test

We can similarly expand the joint test for the 3-2-2-1 indirect effect by testing each of the constituent paths. The test statistics for each path are as follows:

$$t_{a_1} = a_1/\sigma_{a_1} \text{ and } t_{D_{21}} = D_{21}/\sigma_{D_{21}} \text{ and } t_{B^{(2)}} = B^{(2)}/\sigma_{B^{(2)}} \quad (40)$$

For each of these sub-tests, we use a referent t-distribution with degrees of freedom equal to n_3-C-1 where C is the number of school-level predictors in the specific model (Raudenbush & Bryk, 2002; Kenny & Judd, 2014). The power of two-sided tests to detect the indirect effect (i.e., both paths concurrently nonzero) is the product of the power to detect the intervention-mediator path and the power to detect the corresponding mediator-outcome path:

$$\begin{aligned} &P(|t_{a_1}|>t_{critical}\&|t_{D_{21}}|>t_{critical}\&|t_{B^{(2)}}|>t_{critical}) \\ &= (1-t(t_{critical}-t_{a_1}) + t(-t_{critical}-t_{a_1})) \times (1-t(t_{critical}-t_{D_{21}}) \\ &+t(-t_{critical}-t_{D_{21}})) \times (1-t(t_{critical}-t_{B^{(2)}}) + t(-t_{critical}-t_{B^{(2)}})) \end{aligned} \tag{41}$$

where t is the appropriate cumulative t density function and $t_{critical}$ is the corresponding critical value for the appropriate n_3-C-1 degrees of freedom depending on the path model.

Monte Carlo Interval Test

We can further expand the Monte Carlo interval test by drawing plausible values for each path using:

$$\begin{pmatrix} a_1^* \\ D_{21}^* \\ B^{(2)*} \end{pmatrix} \sim t_{n_3-C-1}\left[\begin{pmatrix} \hat{a}_1 \\ \hat{D}_{21} \\ \hat{B}^{(2)} \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2_{\hat{a}_1} & \hat{\sigma}_{\hat{a}_1\hat{D}_{21}} & \hat{\sigma}_{\hat{a}_1\hat{B}^{(2)}} \\ \hat{\sigma}_{\hat{a}_1\hat{D}_{21}} & \hat{\sigma}^2_{\hat{D}_{21}} & \hat{\sigma}_{\hat{D}_{21}\hat{B}^{(2)}} \\ \hat{\sigma}_{\hat{a}_1\hat{B}^{(2)}} & \hat{\sigma}_{\hat{D}_{21}\hat{B}^{(2)}} & \hat{\sigma}^2_{\hat{B}^{(2)}} \end{pmatrix}\right] \tag{42}$$

Inferences regarding the sequential indirect effect are then made on the basis of whether the simulated asymmetric confidence intervals exclude zero. The statistical power of the test is the proportion of asymmetric confidence intervals (e.g., 95%) that exclude zero.

Simulation

We evaluated our 3-2-2-1 power formulas via simulation using the same structure outlined in the 3-2-1 case. The results are summarized in Table 3 (power) and Table 4 (type one error). Similar to the 3-2-1 results, we consistently saw a good correspondence between the predicted and simulated rates for the Monte Carlo and joint tests but only moderate and at times uneven correspondence for the Sobel test. Overall, the Monte Carlo interval test demonstrated the most consistent accuracy and robustness across conditions and is recommended.

Example

Continuing with our earlier 3-2-1 example, let us now introduce instruction as the second or distal mediator. That is, we would like to design a study that intends to use a school-randomized design to probe the impact of an early elementary professional development program in reading (intervention) by examining how it produces changes in teacher knowledge (proximal mediator) which are then passed on to improve instruction (distal mediator), which in turn advances student reading achievement (outcome;

Table 3. Power for 3-2-2-1 mediation.

n_3	n_2	n_1	a_1	a_2	B_1	B_2	D_{21}	ρ_Y^{L3}	ρ_Y^{L2}	$\rho_{M^{(1)}}$	$\rho_{M^{(2)}}$	$R^2_{Y_2^{L3}}$	$R^2_{Y_2^{L2}}$	$R^2_{Y_2^{L1}}$	$R^2_{M_z^{(1)L3}}$	$R^2_{M_z^{(1)L2}}$	$R^2_{M_z^{(2)L3}}$	$R^2_{M_z^{(2)L2}}$	$\widehat{Sobel}$	Sobel	$\widehat{Joint}$	Joint	$\widehat{MC}$	MC
40	4	20	0.29	0.21	0.24	0.33	0.22	0.28	0.15	0.27	0.3	0.38	0.43	0.02	0.14	0.07	0.19	0.11	0.16	0	0.19	0.23	0.17	0.18
60	4	20	0.29	0.22	0.25	0.32	0.22	0.28	0.15	0.27	0.29	0.39	0.43	0.02	0.15	0.07	0.2	0.11	0.23	0.02	0.45	0.49	0.43	0.44
80	4	20	0.29	0.21	0.25	0.33	0.23	0.28	0.15	0.27	0.3	0.39	0.43	0.02	0.15	0.07	0.2	0.11	0.31	0.08	0.68	0.67	0.66	0.67
100	4	20	0.29	0.21	0.25	0.33	0.23	0.28	0.15	0.27	0.3	0.4	0.43	0.02	0.16	0.07	0.21	0.11	0.37	0.17	0.81	0.81	0.81	0.81
60	2	20	0.28	0.21	0.24	0.33	0.22	0.31	0.14	0.27	0.29	0.4	0.43	0.02	0.16	0.07	0.21	0.11	0.19	0.01	0.38	0.38	0.36	0.36
60	4	20	0.29	0.21	0.24	0.33	0.22	0.24	0.12	0.18	0.22	0.48	0.55	0.02	0.23	0.07	0.28	0.1	0.26	0.04	0.52	0.56	0.52	0.52
70	10	10	0.28	0.21	0.24	0.33	0.22	0.26	0.16	0.27	0.3	0.4	0.44	0.02	0.15	0.07	0.19	0.11	0.3	0.08	0.60	0.62	0.58	0.60
80	4	20	0.29	0.21	0.24	0.32	0.22	0.23	0.12	0.18	0.22	0.48	0.56	0.02	0.23	0.06	0.28	0.10	0.34	0.11	0.70	0.72	0.70	0.70
150	10	20	0.29	0.21	0.25	0.33	0.22	0.26	0.15	0.27	0.29	0.41	0.43	0.02	0.16	0.07	0.2	0.11	0.58	0.64	0.95	0.95	0.95	0.95
60	4	20	0.29	0.21	0.16	0.44	0.22	0.29	0.15	0.27	0.3	0.39	0.43	0.02	0.15	0.07	0.19	0.11	0.24	0.04	0.48	0.51	0.47	0.48
50	2	20	0.29	0.22	0.29	0.3	0.15	0.36	0.18	0.06	0.1	0.75	0.87	0.34	0	0.5	0.26	0.55	0.12	0	0.15	0.06	0.14	0.10
50	4	20	0.29	0.22	0.3	0.31	0.15	0.32	0.19	0.08	0.12	0.78	0.87	0.34	0.36	0.52	0.49	0.56	0.15	0	0.21	0.17	0.2	0.17
80	4	20	0.29	0.22	0.3	0.31	0.15	0.32	0.19	0.08	0.12	0.78	0.87	0.34	0.36	0.52	0.49	0.56	0.23	0.03	0.38	0.35	0.39	0.37
80	4	10	0.15	0.11	0.27	0.27	0.15	0.55	0.28	0.34	0.41	0.96	0.97	0.34	0.94	0.63	0.95	0.67	0.14	0	0.26	0.27	0.24	0.23
20	6	20	0.15	0.11	0.28	0.27	0.15	0.54	0.28	0.35	0.41	0.97	0.97	0.34	0.94	0.62	0.95	0.66	0.06	0	0.01	0.01	0.01	0
30	4	10	0.15	0.11	0.27	0.27	0.15	0.55	0.28	0.34	0.41	0.96	0.97	0.34	0.94	0.63	0.95	0.67	0.07	0	0.03	0.03	0.02	0.01
40	6	10	0.28	0.21	0.24	0.33	0.22	0.27	0.15	0.27	0.29	0.39	0.44	0.02	0.15	0.07	0.2	0.11	0.17	0	0.20	0.24	0.18	0.19
40	4	20	0.27	0.2	0.23	0.31	0.22	0.44	0.27	0.35	0.38	0.88	0.83	0.16	0.63	0.41	0.7	0.52	0.17	0.01	0.23	0.25	0.21	0.21
30	4	20	0.23	0.16	0.2	0.29	0.35	0.56	0.3	0.41	0.55	0.96	0.95	0.43	0.87	0.74	0.91	0.79	0.21	0.01	0.27	0.23	0.26	0.22
30	4	20	0.29	0.21	0.22	0.3	0.22	0.29	0.21	0.1	0.09	0.88	0.93	0.33	0.62	0.69	0.64	0.62	0.12	0	0.01	0.05	0.10	0.07
40	4	20	0.29	0.21	0.22	0.29	0.22	0.29	0.21	0.1	0.08	0.88	0.93	0.33	0.63	0.69	0.66	0.62	0.16	0.01	0.17	0.11	0.16	0.15
60	4	20	0.29	0.21	0.22	0.29	0.22	0.29	0.21	0.09	0.08	0.88	0.93	0.33	0.63	0.69	0.65	0.62	0.23	0.05	0.31	0.26	0.33	0.31
80	4	20	0.29	0.21	0.22	0.29	0.22	0.29	0.21	0.09	0.08	0.88	0.93	0.33	0.63	0.69	0.65	0.62	0.30	0.12	0.44	0.37	0.45	0.44
100	4	20	0.29	0.21	0.22	0.29	0.22	0.29	0.21	0.09	0.08	0.88	0.93	0.33	0.62	0.69	0.66	0.62	0.35	0.20	0.53	0.51	0.52	0.55
30	4	20	0.34	0.22	0.42	0.67	0.58	0.09	0.13	0.12	0.11	0.45	0.66	0.59	0.50	0.69	0.59	0.18	0.19	0.08	0.14	0.09	0.14	0.15
40	4	20	0.34	0.22	0.43	0.66	0.58	0.09	0.13	0.12	0.1	0.46	0.66	0.59	0.49	0.69	0.62	0.18	0.24	0.14	0.23	0.18	0.23	0.23
60	4	20	0.34	0.22	0.43	0.66	0.58	0.09	0.13	0.11	0.1	0.46	0.65	0.59	0.49	0.69	0.66	0.17	0.36	0.29	0.40	0.33	0.41	0.41
80	4	20	0.34	0.22	0.43	0.66	0.58	0.09	0.13	0.11	0.1	0.45	0.66	0.59	0.48	0.69	0.67	0.17	0.47	0.45	0.54	0.50	0.56	0.58
100	4	20	0.34	0.22	0.42	0.66	0.58	0.09	0.13	0.11	0.1	0.46	0.66	0.59	0.50	0.69	0.67	0.17	0.57	0.60	0.65	0.63	0.65	0.69
200	4	20	0.34	0.22	0.43	0.66	0.59	0.09	0.13	0.11	0.1	0.46	0.65	0.59	0.49	0.69	0.7	0.17	0.87	0.93	0.93	0.94	0.93	0.95
30	2	10	0.35	0.22	0.37	0.62	0.53	0.12	0.12	0.08	0.14	0.33	0.62	0.59	0.19	0.69	0.42	0.19	0.15	0.03	0.08	0.03	0.08	0.07
40	2	10	0.34	0.23	0.37	0.61	0.55	0.12	0.12	0.07	0.13	0.32	0.63	0.59	0.09	0.69	0.43	0.19	0.21	0.08	0.15	0.07	0.15	0.11
60	2	10	0.35	0.22	0.37	0.6	0.55	0.12	0.12	0.07	0.12	0.33	0.62	0.59	0.06	0.69	0.49	0.18	0.31	0.16	0.32	0.18	0.32	0.28
80	2	10	0.35	0.22	0.37	0.59	0.56	0.12	0.12	0.06	0.11	0.34	0.62	0.59	0.08	0.69	0.53	0.18	0.40	0.27	0.47	0.29	0.46	0.41
100	2	10	0.35	0.22	0.36	0.59	0.55	0.12	0.12	0.06	0.11	0.34	0.62	0.59	0.11	0.69	0.54	0.18	0.48	0.37	0.58	0.44	0.59	0.51
200	2	10	0.35	0.22	0.36	0.59	0.55	0.12	0.12	0.04	0.11	0.34	0.62	0.59	0.39	0.69	0.59	0.18	0.80	0.78	0.90	0.78	0.91	0.83
30	2	5	0.35	0.21	0.37	0.62	0.53	0.12	0.1	0.08	0.14	0.34	0.56	0.59	0.17	0.69	0.42	0.19	0.16	0.03	0.08	0.04	0.09	0.07
40	2	5	0.35	0.21	0.37	0.61	0.55	0.12	0.1	0.07	0.13	0.33	0.56	0.59	0.08	0.7	0.44	0.19	0.21	0.06	0.16	0.07	0.15	0.12
60	2	5	0.35	0.22	0.37	0.6	0.55	0.12	0.1	0.06	0.12	0.34	0.56	0.59	0.03	0.7	0.49	0.19	0.30	0.15	0.31	0.17	0.30	0.26
80	2	5	0.35	0.22	0.36	0.6	0.55	0.12	0.1	0.05	0.11	0.34	0.56	0.59	0.14	0.7	0.52	0.19	0.40	0.27	0.46	0.29	0.45	0.40
100	2	5	0.35	0.21	0.36	0.59	0.55	0.12	0.1	0.05	0.11	0.34	0.55	0.59	0.21	0.7	0.54	0.19	0.48	0.37	0.59	0.45	0.58	0.53
200	2	5	0.35	0.22	0.36	0.59	0.55	0.12	0.1	0.04	0.11	0.34	0.56	0.59	0.6	0.7	0.58	0.18	0.80	0.78	0.91	0.81	0.91	0.84
40	2	5	0.57	0.28	0.32	0.65	0.75	0.39	0.07	0.1	0.26	0.39	0.56	0.59	0.08	0.7	0.53	0.19	0.51	0.39	0.55	0.40	0.56	0.48
60	2	5	0.57	0.29	0.32	0.64	0.74	0.39	0.07	0.09	0.26	0.4	0.55	0.59	0.16	0.7	0.56	0.18	0.70	0.64	0.77	0.64	0.79	0.70
80	2	5	0.58	0.29	0.32	0.64	0.75	0.39	0.07	0.08	0.25	0.39	0.56	0.59	0.27	0.7	0.56	0.18	0.84	0.79	0.90	0.79	0.91	0.85
100	2	5	0.58	0.29	0.32	0.64	0.74	0.39	0.07	0.08	0.25	0.39	0.56	0.59	0.31	0.7	0.57	0.17	0.92	0.88	0.96	0.86	0.96	0.89
30	4	5	0.56	0.3	0.39	0.73	0.81	0.31	0.08	0.14	0.22	0.44	0.57	0.59	0.31	0.71	0.55	0.18	0.44	0.37	0.45	0.40	0.49	0.48
40	4	5	0.56	0.3	0.39	0.72	0.81	0.31	0.08	0.14	0.22	0.45	0.57	0.59	0.3	0.71	0.56	0.18	0.57	0.58	0.61	0.59	0.64	0.65
60	4	5	0.56	0.3	0.39	0.72	0.82	0.31	0.08	0.14	0.22	0.45	0.57	0.59	0.32	0.71	0.57	0.18	0.77	0.83	0.82	0.84	0.84	0.86
80	4	5	0.56	0.3	0.39	0.73	0.81	0.31	0.08	0.14	0.22	0.45	0.57	0.59	0.31	0.71	0.57	0.18	0.88	0.93	0.92	0.94	0.93	0.94
100	4	5	0.56	0.3	0.39	0.73	0.81	0.31	0.08	0.14	0.22	0.45	0.57	0.59	0.31	0.71	0.57	0.18	0.94	0.98	0.97	0.98	0.97	0.98
200	4	5	0.56	0.3	0.39	0.73	0.81	0.31	0.08	0.14	0.22	0.45	0.57	0.59	0.31	0.71	0.57	0.18	1	1	1	1	1	1
500	4	5	0.56	0.3	0.39	0.73	0.81	0.31	0.08	0.14	0.22	0.45	0.57	0.59	0.31	0.71	0.57	0.18	1	1	1	1	1	1
																		Ave	0.40	0.30	0.49	0.45	0.49	0.47

Table 4. Type one error rates for 3-2-2-1 mediation.

n_3	n_2	n_1	a_1	a_2	B_1	B_2	D_{21}	ρ_Y^{L3}	ρ_Y^{L2}	$\rho_{M^{(1)}}$	$\rho_{M^{(2)}}$	$R^2_{Y_2^{L3}}$	$R^2_{Y_2^{L2}}$	$R^2_{Y_2^{L1}}$	$R^2_{M_z^{(1)L3}}$	$R^2_{M_z^{(1)L2}}$	$R^2_{M_z^{(2)L3}}$	$R^2_{M_z^{(2)L2}}$	$\widehat{Sobel}$	Sobel	$\widehat{Joint}$	Joint	$\widehat{MC}$	MC
60	4	20	0	0	0	0	0	0.09	0.1	0.13	0.13	0.48	0.55	0.02	0.33	0.07	0.33	0.07	0	0.01	0	0	0	0
40	6	10	0.29	0.21	0	0	0.22	0.13	0.14	0.27	0.3	0.31	0.39	0.02	0.15	0.07	0.19	0.11	0.05	0	0.01	0.01	0	0.01
40	6	10	0	0	0.24	0.33	0.22	0.24	0.16	0.22	0.24	0.47	0.44	0.02	0.2	0.07	0.28	0.11	0.05	0	0.02	0.01	0.01	0
60	6	20	0.29	0.21	0.25	0	0.22	0.18	0.17	0.27	0.29	0.37	0.43	0.02	0.15	0.07	0.2	0.11	0.03	0	0.03	0.02	0.01	0.01
60	4	20	0	0	0.24	0.33	0.22	0.26	0.15	0.22	0.24	0.45	0.43	0.02	0.2	0.07	0.28	0.11	0.05	0	0.03	0.01	0.01	0.01
60	4	20	0.27	0.21	0.23	0.31	0	0.41	0.27	0.34	0.35	0.89	0.83	0.16	0.63	0.41	0.63	0.41	0.02	0	0.04	0.02	0.02	0.02
80	4	20	0.27	0.21	0.23	0.3	0	0.41	0.27	0.34	0.34	0.89	0.83	0.16	0.63	0.41	0.63	0.41	0.02	0	0.05	0.02	0.02	0.02
60	2	10	0.35	0.24	0	0	0	0	0.04	0.07	0.05	0.12	0.26	0.59	0.03	0.69	0.01	0.11	0.05	0	0	0	0	0
60	2	10	0.35	0.24	0.44	0.63	0	0.08	0.11	0.07	0.05	0.11	0.61	0.59	0.04	0.69	0.02	0.11	0.02	0.01	0.04	0.01	0.02	0.01
60	2	10	0.35	0.21	0.71	0	0.55	0.01	0.07	0.07	0.12	0.28	0.56	0.59	0.04	0.69	0.49	0.19	0.02	0	0.02	0.01	0.01	0.01
																		Ave	0.03	0.00	0.02	0.01	0.001	0.01

Figure 1b). Let us continue with the previous parameter values and further add that we expect the path coefficients between the mediators to be $D_{21} = 0.5$ while that of instruction and achievement is $B^{(2)} = 0.3$. That is, let $a_1 = 0.5$, $a_2 = 0.5$, B_1=0.3, $b_1^{(2)} = 0.1$, $B_2 = 0.3$, $b_2^{(2)} = 0.1$, $D_{21} = 0.3$, $d_{21} = 0.1$, $c' = 0.1$, $\upsilon_Y^2 = \tau_Y^2 = 0.15$, $\sigma_Y^2 = 0.7, \tau_{M^{(1)}}^2 = \tau_{M^{(2)}}^2 = 0.2$, $\sigma_{M^{(1)}}^2 = \sigma_{M^{(2)}}^2 = 0.8, R_{Y_z^{L3}}^2 = R_{Y_z^{L2}}^2 = R_{Y_z^{L1}}^2 = R_{M_{L3_z}^{(1)}}^2 = R_{M_{L2_z}^{(1)}}^2 = R_{M_{L3_z}^{(2)}}^2 = R_{M_{L2_z}^{(2)}}^2 = 0.50$ (e.g., Westine et al., in review). If we sample 20 students per teacher (n_1), 4 teachers per school (n_2), how many schools do we need to yield approximately 80% power to detect the indirect effect? Figure 2b outlines the resulting power functions for the example by detailing the power for the Sobel (long-dash curve), joint (short-dash curves), and Monte Carlo interval tests (solid line curve) for the indirect effect as a function of school-level sample size (n_3). Our application indicates that under the Monte Carlo interval and joint tests roughly 74 schools would yield a power level close to 0.80. For the Sobel test, we would need nearly 200. Under similar conditions, we would need only about 28 schools to have a power level of 0.80 to detect the main effect of 0.445. Supplementary Figure A2 expands these results for plausible ranges of parameter values and suggests that, for example, the required sample size is likely to be somewhat insensitive to misspecifications of the variance explained parameters and somewhat sensitive to misspecifications of the intraclass correlation coefficient parameters.

Discussion

A common charge of many professional development studies is to establish the mechanisms through which program content improves teacher preparation and instruction in ways that advance student learning. This line of inquiry provides a more comprehensive base of evidence because it establishes the extent to which a program shifts primary student outcomes while detailing how the coordinated theory of teacher development comes to (or fails to) scaffold the effects of the program through key intermediate teacher outcomes defining a program's theory of action (Kelcey et al., 2017; Raudenbush & Sadoff, 2008). In this study, we advance this line of inquiry by developing tools to guide the design of such studies.

In education experiments, we often see school-level sample sizes of about 30–100 schools (Spybrook, Shi, & Kelcey, 2016). Our results suggest that with samples of 2 to 4 teachers per school and 10 to 20 students per teacher, school sample sizes within this range may often be sufficient for mediation effects when the study is carefully designed. However, the results also caution that the sample sizes needed to detect mediation effects are not necessarily more or less than the sample size required for the total effect. Differences in the requisite sample size between the mediation and total effects for a given power level are contingent upon the values of other parameters—sometimes the mediation effect will require a larger sample, and under other conditions, the total effect will require a larger sample.

The complex roles and interactions of the parameter values make it difficult to provide simple design strategies. Although parameters associated with the outcome (e.g., outcome variance decomposition, outcome variance explained by covariates) behave intuitively and similar to their counterparts for the total effect, the roles and influence of parameters associated with the mediator (e.g., mediator variance decomposition,

mediator variance explained by covariates) on power were much more complicated and contingent upon the values of concomitant parameters.

The complicated requirements for a causal interpretation of the indirect effects and the complex relationships among parameters governing power emphasize the critical importance of drawing on substantive theory and prior empirical results. A major challenge in researching indirect effects is that they can only be causally interpreted when the design controls for variables that buttress sequential ignorability. For instance, our simulation demonstrated that when a mediator-outcome confounding variable is omitted, the utility of the proposed power analyses is subject to prediction error and is technically dubious. Even with sequential ignorability in place, the efficacy of the proposed power analyses still depends on the maturity of the empirical literature surrounding the targeted intervention, mediator and outcome. In particular, our analyses suggest that the effective use of formulas will depend heavily on the availability of empirical estimates of parameters.

Recent literature has developed an increasingly durable base of design parameter values for planning experiments (e.g., Hedges & Hedberg, 2007; Kelcey & Shen, 2016). To a large extent, however, these studies have predominantly focused on student outcomes and this is a limiting factor in the application of our results. Missing from this literature is, for instance, whether and when we should expect treatment by mediator interaction effects in teacher development studies and the values of design parameters that envelope teacher outcomes. That is, while design parameter values associated with student outcomes are necessary and critical for study design, design parameter values associated with the teacher and other intermediate outcomes become vital if we are to develop effective and efficient studies that systematically build a science of teaching and learning.

Recent research has begun to expand in this direction both in terms of theoretical frameworks detailing key intermediate teacher outcomes (e.g., teacher knowledge, instruction; Desimone, 2009; Phelps et al., 2016) and in terms of empirical estimates (Kelcey & Carlisle, 2013; Kelcey, McGinn, & Hill, 2014; Kelcey & Phelps, 2013b). However, care is needed in the use of the power expressions and subsequent research is needed to expand the scope and scale of these initial efforts. More generally, delineating the sensitivity of power formulas to violations of the assumptions and misspecified parameter values are important areas of research.

To make the results accessible, we have implemented the results in the *PowerUpR* package and in a corresponding *Shiny* application (https://poweruprshiny.shinyapps.io/v104). The *Shiny* application draws on a simple web-based user-interface that requires the user to input prospective parameter values and returns power and sample size estimates for total, mediation, and moderation effects for a wide range of experimental and quasi-experimental study designs.

Funding

This study was supported by grants from the National Science Foundation [1012665, 1760884, 1437745, and 1437692]. The opinions expressed herein are those of the authors and not the funding agencies.

References

Allison, P. (1995). Exact variance of indirect effects in recursive linear models. *Sociological Methodology*, *25*, 253–266. doi:10.2307/271069

Avin, C., Shpitser, I., & Pearl, J. (2005). *Identifiability of path-specific effects*. IJCAI International Joint Conference on Artificial Intelligence (pp. 357–363).

Bauer, D., Preacher, K., & Gil, K. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, *11*(2), 142–163. doi:10.1037/1082-989X.11.2.142

Beasley, T. (2014). Test of mediation: Paradoxical decline in statistical power as a function of mediator collinearity. *The Journal of Experimental Education*, *82*(3), 283–306. doi:10.1080/00220973.2013.813360

Bollen, K. A. (1987). Total, direct, and indirect effects in structural equation models. *Sociological Methodology*, *17*, 37–69. doi:10.2307/271028

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, *33*(8), 3–15. doi:10.3102/0013189X033008003

Brincks, A., Enders, C., Llabre, M., Bulotsky-Shearer, R., Prado, G., & Feaster, D. (2017). Centering predictor variables in three-level contextual models. *Multivariate Behavioral Research*, *52*(2), 149–163. doi:10.1080/00273171.2016.1256753

Carlisle, J., Kelcey, B., Rowan, B., & Phelps, G. (2011). Teachers' knowledge about early reading: Effects on student achievement. *Journal of Research on Educational Effectiveness*, *4*(4), 289–321. doi:10.1080/19345747.2010.539297

Correnti, R., & Rowan, B. (2007). Opening up the black box: Literacy instruction in schools participating in three comprehensive school reform programs. *American Educational Research Journal*, *44*(2), 298–338. doi:10.3102/0002831207302501

Cox, K., & Kelcey, B. (2019). Robustness of statistical power in group-randomized studies of mediation under an optimal sampling framework. *Methodology*, *15*(3), 106–118.

Daniel, R., Stavola, B., Cousens, S., & Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, *71*(1), 1–14. doi:10.1111/biom.12248

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199. doi:10.3102/0013189X08331140

Dong, N., Kelcey, B., & Spybrook, J. (2018). Power analyses for moderator effects in three-level cluster randomized trials. *The Journal of Experimental Education*, *86*(3), 489–514.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. doi:10.1037/1082-989X.12.2.121

Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological Science*, *24*(10), 1918–1927. doi:10.1177/0956797613480187

Hedges, L., & Hedberg, E. (2007). Intraclass correlation values for planning school-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. doi:10.3102/0162373707299706

Hill, H., & Chin, M. (2018). Connections between teachers' knowledge of students, instruction, and student achievement outcomes. *American Educational Research Journal*, *55*(5), 1076–1112. doi:10.3102/0002831218769614

Hong, G., & Nomi, T. (2012). Weighting methods for assessing policy effects mediated by peer change. *Journal of Research on Educational Effectiveness*, *5*(3), 261–289. doi:10.1080/15348431.2012.688421

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*(475), 901–910. doi:10.1198/016214506000000447

Hong, G., & Raudenbush, S. W. (2008). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*, *33*(3), 333–362. doi:10.3102/1076998607307355

Institute of Education Sciences. (2019). *Institute of education sciences request for applications. Education Research Grants CFDA Number: 84.305A*. Retrieved from https://ies.ed.gov/funding/pdf/2019_84305A.pdf

Kelcey, B. (2011). Assessing the effects of teachers' reading knowledge on students' achievement using multilevel propensity score stratification. *Educational Evaluation and Policy Analysis*, *33*(4), 458–482. doi:10.3102/0162373711415262

Kelcey, B., & Carlisle, J. (2013). Learning about teachers' literacy instruction from classroom observations. *Reading Research Quarterly*, *48*(3), 301–317. doi:10.1002/rrq.51

Kelcey, B., Dong, N., Spybrook, J., & Shen, Z. (2017). Experimental power for indirect effects in group-randomized studies with group-level mediators. *Multivariate Behavioral Research*, *52*(6), 699–719.

Kelcey, B., Hill, H., & Chin, M. (2019). Teacher mathematical knowledge, instructional quality, and student outcomes: A multilevel mediation analysis. *School Effectiveness & School Improvement*. *30*(4), 398–431.

Kelcey, B., McGinn, D., & Hill, H. (2014). Approximate measurement invariance in cross-classified rater-mediated assessments. *Frontiers in Psychology*, *5*, 1469. doi:10.3389/fpsyg.2014.01469

Kelcey, B., & Phelps, G. (2013a). Considerations for designing school randomized trials of professional development with teacher knowledge outcomes. *Educational Evaluation and Policy Analysis*, *35*(3), 370–390. doi:10.3102/0162373713482766

Kelcey, B., & Phelps, G. (2013b). Strategies for improving power in school randomized studies of professional development. *Evaluation Review*, *37*(6), 520–554.

Kelcey, B., & Shen, Z. (2016). Multilevel design of school effectiveness studies in sub-Saharan Africa. *School Effectiveness and School Improvement*, *27*(4), 492–510.

Kelcey, B. & Shen, Z. (2019). Strategies for efficient experimental design in studies probing 2-1-1 mediation. *Journal of Experimental Education*, *88*(2), 311–334.

Kelcey, B., Shen, Z., & Spybrook, J. (2016). Intraclass correlation coefficients for designing school randomized trials in education in sub-Saharan Africa. *Evaluation Review*, *40*(6), 500–525.

Kenny, D. A., & Judd, C. M. (2014). Power anomalies in testing mediation. *Psychological Science*, *25*(2), 334–339.

Konstantopoulos, S. (2012). The impact of covariates on statistical power in cluster randomized designs: Which level matters more? *Multivariate Behavioral Research*, *47*(3), 392–420. doi:10.1080/00273171.2012.673898

Kowalski, S., Taylor, J., Askinas, K., Wang, Q., Zhang, Q., & Maddix, W. (2019). Examining empirical evidence for features of effect science teacher professional development. Manuscript submitted for publication.

Kisbu-Sakarya, Y., MacKinnon, D. P., & Miočević, M. (2014). The distribution of the product explains normal theory mediation confidence interval estimation. *Multivariate Behavioral Research*, *49*(3), 261–268. doi:10.1080/00273171.2014.903162

Kreft, I. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*(1), 1–21. doi:10.1207/s15327906mbr3001_1

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, *36*(2), 249–277. doi:10.1207/S15327906MBR3602_06

MacKinnon, D. P., Lockwood, C. M., Brown, C. H., Wang, W., & Hoffman, J. M. (2007). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials: Journal of the Society for Clinical Trials*, *4*(5), 499–513. doi:10.1177/1740774507083434

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83–104. doi:10.1037/1082-989X.7.1.83

Matsumura, L., Garnier, H., & Spybrook, J. (2013). Literacy coaching to improve student reading achievement: A multi-level mediation model. *Learning and Instruction*, *25*, 35–48.

Muthén, B. (2011). *Applications of causally defined direct and indirect effects in mediation analysis using SEM in Mplus*. Retrieved from https://pdfs.semanticscholar.org/e2bd/c70fc0c5f7477580f27b43f632ae62a3f112.pdf

Phelps, G., Kelcey, B., Jones, N., & Liu, S. (2016). Informing estimates of program effects for studies of mathematics professional development using teacher content knowledge outcomes. *Evaluation Review*, *40*(5), 383–409. doi:10.1177/0193841X16665024

Pituch, K. A., Murphy, D., & Tate, R. (2009). Three-level models for indirect effects in school- and class-randomized experiments in education. *The Journal of Experimental Education*, *78*(1), 60–95. doi:10.1080/00220970903224685

Pituch, K. A., & Stapleton, L. M. (2012). Distinguishing between cross-and cluster-level mediation processes in the cluster randomized trial. *Sociological Methods & Research*, *41*(4), 630–670. doi:10.1177/0049124112460380

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*(2), 77–98. doi:10.1080/19312458.2012.679848

Qin, X., & Hong, G. (2017). A weighting method for assessing between-site heterogeneity in causal mediation mechanism. *Journal of Educational and Behavioral Statistics*, *42*(3), 308–340. doi:10.3102/1076998617694879

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173–185. doi:10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. SAGE.

Raudenbush, S., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*(1), 5–29. doi:10.3102/0162373707299460

Raudenbush, S., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness*, *1*(2), 138–154. doi:10.1080/19345740801982104

Rhoads, C. (2011). The implications of "contamination" for experimental design in education. *Journal of Educational and Behavioral Statistics*, *36*(1), 76–104.

Roth, K. J., Wilson, C. D., Taylor, J. A., Stuhlsatz, M. A. M., & Hvidsten, C. (2019). Comparing the effects of analysis-of-practice and content-based professional development on teacher and student outcomes in science. *American Educational Research Journal*, *56*(4), 1217–1253. doi:10.3102/0002831218814759

Scher, L., & O'Reilly, F. (2009). Professional development for K–12 math and science teachers: What do we really know? *Journal of Research on Educational Effectiveness*, *2*(3), 209–249.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*(1982), 290–312. doi:10.2307/270723

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. *International Journal of Research & Method in Education*, *39*(3), 255–267.

Steen, J., Loeys, T., Moerkerke, B., & Vansteelandt, S. (2017). Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology*, *186*(2), 184–193. doi:10.1093/aje/kwx051

Taylor, A. B., MacKinnon, D. P., & Tein, J. Y. (2008). Tests of the three-path mediated effect. *Organizational Research Methods*, *11*(2), 241–269. doi:10.1177/1094428107300344

Usami, S. (2019). Confidence interval-based sample size determination formulas and some mathematical properties for hierarchical data. *British Journal of Mathematical and Statistical Psychology*. Advance online publication. doi:10.1111/bmsp.12181

VanderWeele, T. J. (2008). Ignorability and stability assumptions in neighborhood effects research. *Statistics in Medicine*, *27*(11), 1934–1943. doi:10.1002/sim.3139

VanderWeele, T. J. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods & Research*, *38*(4), 515–544. doi:10.1177/0049124110366236

VanderWeele, T., Hong, G., Jones, S., & Brown, J. (2013). Mediation and spillover effects in group-randomized trials: A case study of the 4Rs educational intervention. *Journal of the American Statistical Association*, *108*(502), 469–482. doi:10.1080/01621459.2013.779832

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, *2*(4), 457–468. doi:10.4310/SII.2009.v2.n4.a7

Westine, C., Unlu, F., Taylor, J., Spybrook, J., Zhang, Q., & Anderson, B. (in review). Design parameter for impact evaluation of science and mathematics intervention involving teacher outcomes. Unpublished manuscript.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher PD affects student achievement*. Washington, DC: U.S. Department of Education.

Zhang, Z., Zyphur, M., & Preacher, K. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods*, *12*(4), 695–719. doi:10.1177/1094428108327450